Medical Diagnosis based Bio-inspired Artificial Immune System (AIS)

Author

MUHAMMAD WASEEM TAHIR

00000206164

Supervisor

BRIG DR. JAVAID IQBAL

DEPARTMENT OF MECHATRONICS ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

ISLAMABAD

JULY, 2021

Medical Diagnosis based Bio-inspired Artificial Immune system (AIS)

Author

MUHAMMAD WASEEM TAHIR

00000206164

A thesis submitted in partial fulfillment of the requirements for the degree of

MS Mechatronics Engineering

Thesis Supervisor:

BRIG DR. JAVAID IQBAL

Thesis Supervisor' signature:____

DEPARTMENT OF MECHATRONICS ENGINEERING

COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY,

ISLAMABAD

JULY, 2021

Declaration

I certify that this research work titled "*Medical Diagnosis based Bio-inspired Artificial Immune System (AIS)"* is my own work. The work has not been presented elsewhere for assessment. The material that has been used from other sources it has been properly acknowledged / referred.

Signature of Student
MUHAMMAD WASEEM TAHIR
00000206164

Language Correctness Certificate

This thesis has been read by an English expert and is free of typing, syntax, semantic, grammatical and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student
MUHAMMAD WASEEM TAHIR
00000206164

Copyright Statement

- Copyright in text of this thesis rests with the student author. Copies (by any process) either in full, or of extracts, may be made only in accordance with instructions given by the author and lodged in the Library of NUST College of E&ME. Details may be obtained by the Librarian. This page must form part of any such copies made. Further copies (by any process) may not be made without the permission (in writing) of the author.
- The ownership of any intellectual property rights which may be described in this thesis is vested in NUST College of E&ME, subject to any prior agreement to the contrary, and may not be made available for use by third parties without the written permission of the College of E&ME, which will prescribe the terms and conditions of any such agreement.
- Further information on the conditions under which disclosures and exploitation may take place is available from the Library of NUST College of E&ME, Rawalpindi.

Acknowledgements

“In the Name of Allah, the Most Beneficent, the Most Merciful. All the praises and thanks be to Allah, the Lord of the worlds” (Quran 1:1-2)

First of all, I would like to thank my Creator **ALLAH ALMIGHTY** for providing me adequate strength, courage and aptitude to accomplish such a gigantic task and to guide me throughout this work. I couldn’t have done anything without **HIS** myriad Guidance and Blessings. Whosoever has helped me throughout the course of my thesis, was through **HIS** ultimate Wish and Will. **HE** is the **ONE** we worship and seek for Help. No doubt, no one is worthy of praise but **ALLAH ALMIGHTY**.

Special thanks to my incomparable & beloved Parents who not only brought me up once I was unable to walk, but also remained my ultimate support and source of motivation, and continued nonstop guidance throughout my life. *"My Lord, have mercy upon them as they did bring me up when I was young" (Quran 17:24).*

My foremost and sincere thankfulness to my supervisor, Professor Dr. Javaid Iqbal for his supervision, fortitude and continuous support throughout this journey. Besides the academic support his guidance provided me power to believe in myself, bring out my best, value my parents and pray none other than **ALLAH ALMIGHTY** for success. Owing to his competent supervision, I have learnt that persistence hard work and belief in **ALLAH ALMIGHTY** are the foundations to success in life here and hereafter.

I would like to express my sincere gratitude to Brig Dr. Nasir Rashid for untiring help, ceaseless persistence, incessant support and endless guidance throughout my thesis work. Each time I got stuck, he was always there for me with a viable solution. Without his undue assistance, it wouldn’t have been possible for me to complete my research work at all. Here, I would also like to thank Dr. Mohsin Islam Tiwana for his kind guidance throughout my thesis and degree, being member of my thesis guidance and evaluation committee. I would also like to pay my gratitude to Dr. Umar Shahbaz, Dr. Amir Hamza and Dr. Mubashir Saleem for tremendous support and cooperation

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

*This work is solely dedicated to my **extraordinary parents** (M. A. Tahir and Misbah Tahir) who have always been a source of consistent support and motivation for me, my lovely **siblings** (Waqas, Afshan and A Aleem), my dearest friend **Hamid**, my adorable **kids** (Hashir & Hareem) and my beloved **spouse (**Mariam**),** without whom, I would have never been able to achieve such a delightful milestone in my entire life.*

Abstract

Remedial disease diagnoses and timely provision of apt treatment always remained a challenge for healthcare professionals. Ongoing depletion of colossal healthcare data, can be converted to into valuable information, with reduced cost, wee processing time and improved diagnosis for life-threating diseases, such as heart diseases. As per World Health Organization (WHO), heart related diseases are the prime reason of mortality, representing 29% of deaths in Pakistan [1] and 31% of the global deaths [2]. On an average, 247 deaths per million, have been recorded and mostly due to delayed diagnosis in Pakistan. Being a major universal health concern, heart disease, was selected for medical diagnosis.

Biological Human Immune System (BHIS) is a complex, robust and an adaptive system that defends human body from foreign or unknown pathogens; capable of self-monitoring, executing optimum disease detection, and active learning from data in memory cells. It comprises of group of molecules, which distinguish cells self (S) or non-self (NS), within the body. Several concepts from the BHIS are applied for real-world glitches using Artificial Immune System (AIS), which are ability to learn and memory to store. Bio-inspired techniques, are being employed in intellegent computing, security, and engineering. However, relatively little research focused on bio-inspired AIS based classifiers in machine learning for the prediction of the presence of heart disease.

With headway in data science, benefits for healthcare industry are being realized across the globe. The key goal is to develop a diagnosis system for precisely envisaging existence of heart disease, through a novel Bio-Inspired Classification (BIC) based on Negative Selection Artificial Immune System (NSAIS), optimized through GA, with UCI Cleveland Dataset. Experimental outcomes show that bio-inspired algorithm outperformed six classical ML methods. This research not only aids in utilizing available medical data, but also enables the doctors to intelligently diagnose heart disease.

Keywords: *Heart attack, heart disease prediction, classification algorithms, medical diagnosis, data mining, machine learning tools, negative selection algorithm (NSA), Genetic Algorithm (GA), Bio-inspired classification (BIC), artificial immune system (AIS)*

Table of Contents

Abstract .... **viii**

Table of Contents .... **ix**

List of Figures .... **xiii**

List of Tables .... **xv**

Nomenclature .... **xvi**

Chapter 1 .... **1**

INTRODUCTION **1**

1.1 Motivation 1

1.1.1 Heart Disease Mortality 2

1.1.2 Medical Data / Clinical Decision Support (CDS) 2

1.1.3 Medical Data Mining 3

1.1.4 Bio-inspired Medical Diagnosis 4

1.2 Research Objectives & Rationale 4

1.2.1 Problem Definition 4

1.2.2 Research Objectives 4

1.3 Thesis Organization 5

Chapter 2 .... **7**

BACKGROUND STUDIES **7**

2.1 Heart Disease 7

2.1.1 Heart Anatomy 8

2.1.2 Risk Factors 9

2.1.3 Symptoms of Heart Disease 10

2.1.4 Heart Disease Diagnosis 10

2.2 Machine Learning in Health Care 11

2.2.1 Heart Disease Dataset 11

2.2.2 Traditional Data Mining Techniques 13

2.2.3 Bio-Inspired Data Mining Techniques 26

2.2.4 Performance Measure of a Disease Prediction System 30

2.2.5 Bench Mark Research 30

2.2.6 Gap Identification 31

Chapter 3 34

THE IMMUNE SYSTEM 34

3.1 Biological Human Immune System (BHIS) 34

3.1.1 Physiology of Immune System 36

3.1.2 Thymic Negative Selection – self (S) and non-self (NS) discrimination 38

3.2 Artificial Immune System (AIS) 40

3.2.1 Negative Selection Artificial Immune System (NSAIS) 41

3.3 Genetic Algorithm (GA) 44

3.4 Applications of NSAIS & GA 45

Chapter 4 47

PROPOSED METHODOLOGY 47

4.1 An Overview of Proposed Methodology 47

4.1.1 Description of Data Set 47

4.1.2 Data Pre-processing 48

4.1.3 Validation Techniques 48

4.2 Mode 1-Traditional ML Classification........ 48

4.2.1 Dimensionality Reduction through Principal Component Analysis (PCA)........ 49

4.2.2 Training Classification Model with Traditional ML Techniques 49

4.2.3 Heart Disease Prediction Through Trained traditional ML Model........ 50

4.3 Mode 2 – Proposed Bio-inspired Classification (BIC) based on NSAIS........ 50

4.3.1 Random Detector Generation through NSAIS........ 52

4.3.2 Optimization of detectors sets using Genetic Algorithm (GA).........54

4.3.3 Training Model with Proposed BIC based on NSAIS Classifier........ 60

4.3.4 Heart Disease Prediction through BIC based on NSAIS 62

Chapter 5 63

RESULTS 63

5.1 Heart Disease Prediction 63

5.1.1 Accuracy Results with Traditional ML Algorithms (Mode 1)........ 64

5.1.2 Accuracy Results with Proposed BIC based on NSAIS (Mode 2) 64

5.2 Analysis and Discussion........ 65

5.2.1 Classification Results in different Data Mining Tools........ 65

5.2.2 Comparison with Various Benchmark Researches........ 66

5.2.3 Traditional ML Algorithms vs BIC based on NSAIS........ 67

5.2.4 Performance Analysis: BIC based on NSAIS optimized with GA........ 68

5.2.5 Discussion on Results 68

5.3 Validation of the Results 69

5.3.1 Description – Miscellaneous Datasets 69

5.3.2 Validation Results – Miscellaneous Datasets 69

5.3.3 Analysis of Results - Miscellaneous Datasets.. 71

Chapter 6 72

CONCLUSIONS AND RECOMMENDATIONS.. 72

6.1 Contribution.. 72

6.2 Conclusions 73

6.3 Recommendations and Future Works 74

REFERENCES.. 75

List of Figures

Figure 2.1. Fats Deposited in Arteries [12].. 8

Figure 2.2. Heart anatomy [13]... 9

Figure 2.3. Condition during Heart Disease [12]... 10

Figure 2.4. SVM classifier– uses the optimal hyperplane for generalization....................... 15

Figure 2.5. Fluctuations in prediction accuracies [31].. 18

Figure 2.6. A typical classification process in prediction of heart disease........................... 32

Figure 3.1. BHIS multilayer structure.. 36

Figure 3.2. Lymphocytes with surface molecules 37

Figure 3.3.BHIS - pattern recognition 38

Figure 3.4. Bone marrow and Thymus [77]... 38

Figure 3.5. Thymic negative selection and maturation process of T-cells 39

Figure 3.6. Negative Selection of Detectors [65] 41

Figure 3.7. Self-Nonself region in Vector Space.. 43

Figure 3.8. Matured Detectors. 43

Figure 3.9. Discarded Detectors.. 44

Figure 3.10. Biological Chromosome vs Computational chromosome [82]. 45

Figure 4.1. Traditional ML Methodology.. 50

Figure 4.2. Impersonation between Thymic Negative Selection BHIS and NSAIS 51

Figure 4.3. Detectors with void in non-self region 52

Figure 4.4. Optimized Detectors in NS region 53

Figure 4.5. Detector generation for NSAIS 53

Figure 4.6. GA optimized negative selection detector optimization 54

Figure 4.7. Variable sized NS detectors with a S sample in Vector Space............................ 57

Figure 4.8. Flow diagram of proposed BIC based on NSAIS 61

Figure 5.1. Classification accuracy for BIC based on NSAIS.. 65

Figure 5.2. Classification Accuracies in two ML Platforms.. 66

Figure 5.3.Accuracy with Proposed & Benchmark Approaches 67

Figure 5.4. Accuracy with Traditional ML Algos vs BIC based on NSAIS 68

Figure 5.5 Percentage classification for Miscellaneous Validation Datasets 70

List of Tables

Table 2.1 Cleveland Dataset attributes 11

Table 2.2 Status of 13 Independent attributes.. 13

Table 2.3 Classfication accuracy for heart disease with Traditional ML Approaches. 24

Table 2.4 Classfication accuracy for heart disease with Bio-inspired techniques. 29

Table 2.5 Binary Classification accuracy of bench mark approach....................................... 31

Table 4.1 Detection accuracy of Detector A(Class 0) 59

Table 4.2 Detection accuracy of Detector B(Class 1).. 59

Table 4.3 Max detection accuracy of the Detectors... 60

Table 4.4 Percentage training detection accuracies of selected detector sets 60

Table 5.1 Classification accuracy (%) with Traditional ML Algorithms 64

Table 5.2 Percentage classification accuracy with BIC based on NSAIS 64

Table 5.3 Comparison of results of proposed approaches with bench mark research............ 66

Table 5.4 Maximum detection accuracy.. 69

Table 5.5 Percentage training detection accuracies 70

Table 5.6 Percentage classification accuracy NSAIS 70

Nomenclature

Acronyms

AIS	Artificial Immune System
AIRS	Artificial Recognition System
APCs	Antigen Presenting Cells
AB	Antibody
AG	Antigen
BHIS	Biological Human Immune System
BCR	B-cell Receptors
ECG	Electrocardiogram
GA	Genetic Algorithm
k-NN	k – Nearest Neighbour
LDA	Linear Discriminant Analysis
MLP	Multilayer Perceptron
NSAIS	Negative Selection Classification Algorithm
NSAIS	Negative Selection Artificial Immune System
NSA	Negative Selection Algorithm

NN	Neural Networks
RNSA	Real Valued Negative Selection Algorithm
SVM	Support Vector Machines
TCR	T-cell Receptors

Symbols

Dn	Number of detectors
D_{n-max}	Maximum number of detectors
d^j	Center point of jth detector
$dist$	Euclidean distance
f_j	Initial fitness value
f_j^*	Final fitness value
μ	Mean value
N	Number of self samples in set
r^j	Radius of jth detector
R^k	Radius of kth self-sample
R^{Self}	Radius of self-detector
$s(n)$	Raw signal
s_j	Diversity factor
Ω	Set of self-samples

Ω^i	ith self-sample (training sample)
ξ_{ij}	Distance based similarity between candidate and selected detector
x	Randomly generated detector set

Chapter 1

INTRODUCTION

The first chapter of this work includes introduction and motivation to develop a disease diagnosis system using data mining techniques. *The basic requirement is* ***heart disease prediction through Bio-inspired Classification (BIC)*** *based on* ***Negative Selection Artificial Immune System (NSAIS),*** *along with problem definition, and the research objective.* The planned outline of thesis is discussed towards the end of this chapter.

1.1 Motivation

Keeping in view the modern technology and specially the human immune system, machine learning /data mining has emerged as an important research area. Data is presented to machine learning algorithms for training, validation and testing. Generally, medical professionals in heart are able to predict chance of heart attack with max of 67% accuracy [3]. In addition, the current COVID-19 epidemic conditions, also necessitates the doctors an intellegent support system for aiding in the precise heart diseease prediction. To improve robustness of medical data by experimenting with a human immune-inspired classification algorithm, this work is aimed achieving higher classification accuracy for heart disease prediction using

- Traditional Machine Learning (ML) Algos
- Bio-inspired Classification (BIC) Algos

1.1.1 Heart Disease Mortality

Heart diseases are one of the prime causes of high mortality rate in the world, representing 31% (17.9 million) of global deaths in 2016 [2]. Out of which, 85% death were due to heart attack or stroke. According to American Heart Association, globally 18.6 million [4] deaths attributed to CVD in 2019, representing an increase of 17.1% from 2010 [4]. Heart disease occur due to disruption of blood flow in the body. Responsible factors include narrowing/blockage/deposition of fats of blood vessels supplying blood to heart, excessive unhealthy food, tobacco utilization, extreme physical latency, having extra body fat or being overweight, and disproportionate use of liquor.

1.1.2 Medical Data / Clinical Decision Support (CDS)

Personalized clinical predictions play a vital role in healthcare and leads to improved disease prevention, through clinical prognosis of patient's health. Heart specialists usually adapt a cerebral prediction method based on former knowledge or past experience, research, & interaction with patients. In 2012, worldwide healthcare data was around $5x10^5$ Terabytes, which was estimated to reach approx. $2.5x10^7$ Terabytes in 2020 [5]. According to Dell EMC, health data has grown up with the rate of 878 %, since 2016 [6], reaching 8.41 petabytes (PB) on average in 2018. With the inundation of health data, effective clinical prediction methods were required to address such challenges.

Clinical Decision Support (CDS) provides timely information, which results in good-quality health care. It is a sophisticated method developed to assist healthcare professionals for clinical tasks like diagnosis, by amalgamating knowledge and data to provide timely decision support. It includes predictive tools and databases that can provide patients' basic info, reminders for preventive care, personalized medicine and creating alerts and thus acts as second opinion establishing effective diagnosis. In addition, few decision support systems utilize available hospital data along with artificial intelligence means [7], for detecting critical conditions, precisely predicting the presence of diseases and results in enhanced health outcomes, evasion of errors & better efficiency, cost-benefit, and patient satisfaction.

1.1.3 Medical Data Mining

Data mining obtains knowledge and valueable information from large set of available data. Nowadays, datamining is seen in the fields of education, data analytics, healthcare, electronic media, economic studies, bio-informatics and weather analysis. In healthcare, it has become most efficient resource as it can overcome depletion of massive healthcare data, predicting diseases, and assisting healthcare professionals in accurate decision making [8].

Medical Data mining is extraction of valuable information by examining huge raw data which is previously unknown, but quite significant. It is used to ascertain decisions, estimate and predict using different algorithms. It allows doctors to see which attributes are more significant for diagnosis. Data mining tools can compare causes, associate symptoms, infer appropriate treatment, track the chronic diseases and identify high-risk patients. Medcial data mining can help improve patient satisfaction, decrease treatment costs and provide high-quality care. Thus, by using medical data mining it is able to get an insight on a patient's history and is able to provide necessary clinical support.

1.1.4 Bio-inspired Medical Diagnosis

A bio-inspired/biomemetic approach encompasses mimicking biological nature to deal with the real-world problems. These techniques have been widely used in fields of intellegent computing (fault and anoamly or virus detection), security (intrusion detection), data analysis (unattended clustering) and engineering (optimization, robotics) [9]. Likewise, these concepts can be translated into the fields of medical diagnosis. Being a sensitive task, medical diagnosis needs to be executed precisely to ensure accurate results. Nevertheless, very little work has been done the field of disease diagnosis thorugh the bio-inspired techniques. Hence, this research focuses on the bio-inspired technique to clasify, optimize and thus predict the presence of the heart disease.

1.2 Research Objectives & Rationale

1.2.1 Problem Definition

The challenge is to identify presence of heart disease with highest accuracy to assist healthcare professionals in medical diagnosis. Since the heart disease is the primary cause of death and colossal healthcare data is depleted, achieving highest classification accuracy in patient-oriented prediction models always remains a challenge for the researcher.

“An efficient method is required to achieve a high accuracy for correctly predicting presence of heart disease to avoid large heart failure rates”

1.2.2 Research Objectives

The goal of this study is to develop a method to correctly classify the heart disease prediction. The main objectives of the research are:

- Carry out study of existing methods for classification of available medical data for disease diagnosis.
- Using smart bio-inspired technique to develop an improved method of classification.
- Evaluate performance of proposed bio-inspired classification algorithm in comparison to the benchmark techniques.
- Develop an intelligent classification algorithm, capable of classifying classify, optimize and thus predict the presence of the heart.

1.3 Thesis Organization

This thesis is organized in six chapters; outline of the organization is as under:

- Literature review has been included in second chapter with background studies of related to heart diseases, risk factors, and anatomy of heart along with various classification methods. Detailed literature review with the relevant traditional machine learning methods being used in healthcare along with acquaintance with bio inspired classification methodologies is presented.
- Third chapter of the thesis describes the concept and principles of Biological Human Immune System (BHIS) theory followed by bio inspired Artificial Immune System (AIS). AIS is further explained in detail with the help of framework to understand and correlate AIS with BHIS.
- Fourth chapter comprises of the overview of Proposed Methodology related to the Classification of heart disease prediction using traditional ML techniques & Bio-

inspired Classification (BIC) technique based NSAIS, optimized through Genetic Algorithm (GA).

- Fifth chapter presents results of proposed methodology in comparison with traditional ML techs and the benchmark research. Moreover, Predictive analysis on the same dataset as of benchmark research and Miscellaneous dataset were used to validate the performance of our proposed system.
- Main contributions, recommendations and future work are part of the last chapter of the thesis.

Chapter 2

BACKGROUND STUDIES

This chapter presents an overview of human heart along with background studies related to heart disease diagnosis through various machine learning techniques. Several studies have been conducted for calculating prediction accuracies using different classifiers for effictive diagnosis of heart illness. At the end of the chapter the gap is identified in the field of research basing on the analysis.

2.1 Heart Disease

Heart disease is the leading cause of deaths worldwide . Remedial diagnosis of heart disease is an intricate task for medical practitioners, as it not only involves ample medical examinations, but also requires a lot of experience to diagnose the heart disease. Different types of heart diseases include coronary, congenital, rheumatic, myocarditis, arrhinia and angina. These can occur as early as at the age of 18 years and are only detected when the 70% damage has taken place [10]. For an effective heart disease diagnoses, a little acquaintance with human heart and its working is essential.

Heart is a muscular organ with size of a human fist. It pumps oxygenated blood through arteries and veins. Average male heart weights 280 g to 340 g (10 to 12 ounces) and in contrast, female heart is around 230 g to 280 g (8 to 10 ounces). It beats 100,000 times and pumps 2,000 gallons a day. Henceforth, heart of an average 80-year-old beats around 2.5 billion times and pumps around 60 million gallons blood in lifetime. 60,000 miles of blood vessels for delivery of blood are present in a normal human body [11]. With inadequate blood flow, heart muscle cannot get essential nutrients and oxygen to function. Heart disease occurs when fat, and cholesterol are deposited inside the arteries and blood vessels leading to the heart, as shown in figure 2.1.

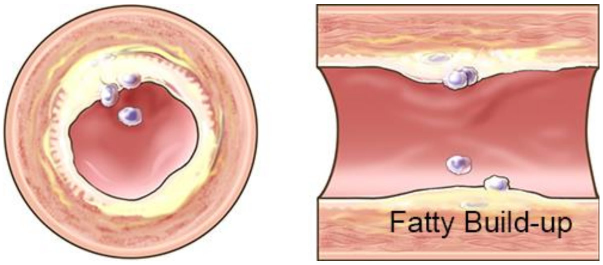


Figure 2.1. Fats Deposited in Arteries **[12]**

2.1.1 Heart Anatomy

Heart pumps blood to all of the body's organs, tissues and cells and also delivers oxygen along with it. In the process, it also exhales carbon dioxide and waste products. Anatomically, the heart is made up of four separate chambers with muscular walls: two upper chambers; left atrium (LA) and right atrium (RA), and two lower chambers; left ventricle (LV) and right ventricle (RV). RA receives non-oxygenated blood from body and pumps it to RV, which pumps blood through the pulmonary artery (PA) to the lungs, where it again becomes oxygenated. LA receives the oxygenated blood from lungs and then pumps it to LV, which sequentially pumps this oxygen-rich blood through the aortic valve to the aorta and the rest of the body. This anatomy of the heart is as illustrated in figure 2.2.

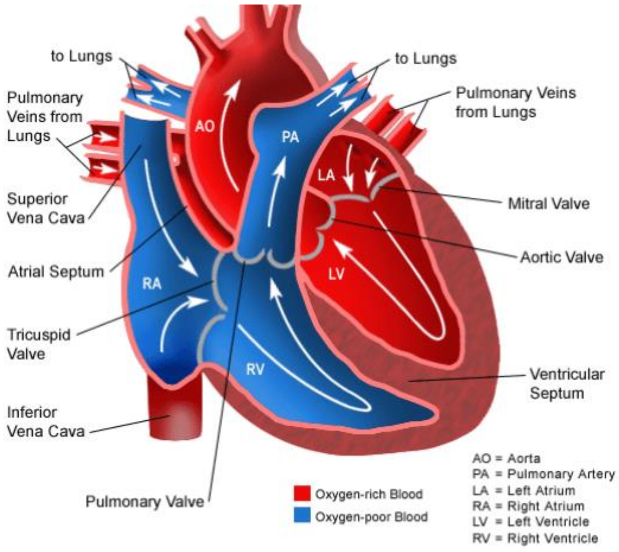


Figure 2.2. Heart anatomy **[13]**

2.1.2 Risk Factors

A risk factor is something that intensifies the chance a disease. They can be classified as non-modifiable i.e., nothing can be done and they keep causing heart disease (e.g., age, gender, family history, and ethnicity) and modifiable i.e., something can be done to cater them (e.g., foods, habits, and stress). Misc. modifiable risk factors include smoking, unhealthy diet, high carbohydrate diet, obesity, poor sleep hygiene, inactivity, hypertension or blood pressure, high levels of the cholesterol, and diabetes According to Interheart study for heart attacks, which scrutinized risk factors in 52 countries, people consuming the diet comprising of fried foods, salty snacks and excessive meat had a 35% greater heart attack risk, as compared to those people not consuming such foods. Moreover, people who consumed fruits and vegetables in their foods had 30 % less risk of heart attack in comparison to people who didn't consume fruits and vegetables in their food [11].

2.1.3 Symptoms of Heart Disease

Common symptoms include any sort of chest discomfort or pain, feeling of shortness of breath, arm pain, heart burn, nausea or vomiting, stomach pain, fatigue and "cold" sweat. If symptoms last longer than 5 minutes, emergency medication is required. If symptoms are timely recognized, it can help in prompting requisite medical treatment and hence save lives. Figure 2.3 shows the area of applications of these common symptoms found during the heart disease.

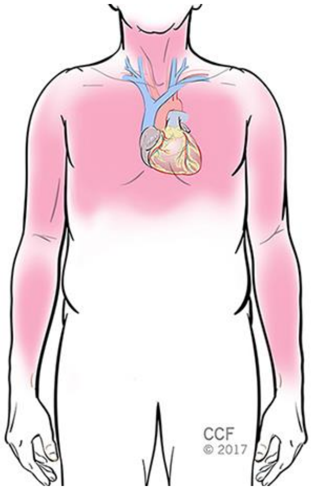


Figure 2.3. Condition during Heart Disease **[12]**

2.1.4 Heart Disease Diagnosis

Reducing use of salt & fried things, eating healthy fruits and vegetables, regular physical activity can reduce the risk behind heart disease. Apart from it, 80% heart diseases can be prevented by healthy behaviours. Heart disease can be addressed only by early detection or prognosis of the risk of disease. Timely diagnosis of heart disease is important for prevention of heart failure. Latest non-intrusive methods such as machine learning can prove to be quite effective in predicting heart disease presence in advance.

2.2 Machine Learning in Health Care

Machine Learning is a subset of Artificial Intelligence(AI), which is the ability of systems to learn thing without being properly programmed. Health care data mining has made it possible to integrate availabe medical resources by exploring untamed patterns in medical data, with machine learning classification techniques andhence for the prediction of diseases. Numerous data mining or machine learning techniques are being utilized to diagnose miscelleneous ailments. These include Diabates [14], Hepatitus [15], Acute Appendicitis [16], Skin Disease [17], Epilepsy [18], Cardiovascular/Heart diseases [19], Breast Cancer [20], Liver Disorders [21], Hemorrhagic Stroke [22] & Childhood leukemia [23]. However, heart disease has been nominated for medical diagnosis through machine learning techniques, being the reason of highest mortality rate in the world in general and Pakistan in specific.

2.2.1 Heart Disease Dataset

Out of multiple medical data sources, Heart Disease Data repository [24] is an extensively used open-source dataset, which contains databases from four hospital i.e., Cleveland, Switzerland, Hungarian, and Long-Beach. In particular, Cleveland dataset is widely used for exploring ML concepts. It has 303 instances and 76 attributes. 14 Attributes used for the prediction of the heart disease presence are described in the Table 2.1

Table 2.1 Cleveland Dataset attributes

Ser	**Attribute**	**Description**
1	Age	Displays the age of the individual.
2	Sex:	Displays the gender of the individual using the following format: 1 = male, 0 = female
3	Chest-pain type: Cp	Displays the type of chest-pain experienced by the individual using the following format: 1 = typical angina 2 = atypical angina

		3 = non — anginal pain 4 = asymptotic
4	Resting Blood Pressure: BP	Displays the resting blood pressure value of an individual in mmHg (unit) for the patient while admitted to the hospital. [Minimum BP: 94, Maximum BP: 200]
5	Serum Cholesterol: Chol	Displays the serum cholesterol in mg/dl (unit). [Minimum Chol: 126, Maximum Chol: 564]
6	Fasting Blood Sugars	compares the fasting blood sugar value of an individual with 120mg/dl. If fasting blood sugar > 120mg/dl then: 1 (true) else: 0 (false)
7	Resting ECG: restecg	displays resting electrocardiographic results 0 = normal 1 = having ST-T wave abnormality 2 = left ventricular hypertrophy
8	Max heart rate achieved:	displays the max heart rate achieved by an individual.
9	*Exercise induced angina* Exang:	1 = yes 0 = no
10	*ST depression induced by exercise relative to rest*: old peak	Displays the value which is an integer or float
11	*Peak exercise ST segment*: Slope	1 = upsloping 2 = flat 3 = Down sloping
12	*Number of major vessels (0–3) colored by fluoroscopy*:	showing that how many vessels are colored.
13	*Thal*:	Shows chest pain or breathing difficulty. Four values showing the result of Thallium test 3 = normal 6 = fixed defect 7 = reversible defect
14	Diagnosis of heart disease:	Class column or Label column Displays whether the individual is suffering from heart disease or not: 0 = absence 1 = present.

As far as the class distribution is concerned, 164 out of 303 patients do not have the heart disease, while the rest of 139 patients have a heart disease. As far as the class distribution is concerned, 164 out of 303 patients do not have the heart disease (Class 0), while the rest of 139 patients have a heart disease (Class 1). Nevertheless, only 14 attributes selected for heart disease prediction; one dependent variable (Class label), and 13 independent variables. Status of 13 independent variables is tabulated in Table 2.2.

Table 2.2 Status of 13 Independent attributes

Type	No of Attributes	Description
Nominal	3	cp, restecg, thal,
Real	6	age, trestbps, col, max heart rate, old peak, ca
Binary	3	sex, fbs, exang

2.2.2 Traditional Data Mining Techniques

Numerous data mining methods have been studied for efficient prediction of heart disease problems. Classification algorithms aid in learning the relationship between of attributes and class label [7]. Most reserchers applied various data mining tools for prediction of the heart disease and found out classification accuracies using different ML algorithms, such as Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Naïve Bayes, Logistic Regression, Random Forest, Decision Tree etc.

2.2.2.1 Data Mining Tools

In order improve medical diagnosis, ample data mining tools are being utilized by the researchers for machine learning studies [25] i.e., MATLAB, Orange, Weka, RapidMiner, Knime, and Scikit-learn. Brief overview of each data mining tool is presented.

2.2.2.1.1 MATLAB

MATLAB is a programming language, which is developed by Math Works. It offers to perform interactive classification with the use of its Classification Learner app, which can perform supervised learning tasks e.g., exploring data, feature selection, validation, training models, and assessing results. It can classify data using various algorithms and compare the results in the same environment.

2.2.2.1.2 Weka

WEKA is open-source machine learning software accessed through a graphical user interface. It builds ML pipelines, train the classifiers, and run requisite evaluations without coding.

2.2.2.1.3 Orange

Orange is an open-source software used for ML and data mining, developed by Bioinformatics Laboratory, which builds data analysis workflows, with a large, toolbox.

2.2.2.1.4 RapidMiner

RapidMiner is an Open source & extensible tool, which provides a data science platform, which chooses responses based on users' preferences. It is used for data prep, machine learning, and model deployment and can be used for Offline and online analysis.

2.2.2.1.5 Knime

KNIME Analytics Platform is the open-source data science software. Being an intuitive software, it makes understanding of the data and designing data science workflows accessible to everyone.

2.2.2.1.6 Scikit-Learn

Scikit-learn is an open-source ML library, which performs supervised and unsupervised learning. It also provides tools for data preprocessing, selecting models and evaluation.

2.2.2.2 Traditional ML Algorithms

2.2.2.2.1 Support Vector Machine (SVM)

Support vector machine algorithm works best where the data has binary classes. It finds a best hyper plane that separates data points of one class from another class. Hyper plane is the separation between the two critical points or members (also called support vectors). A linear-SVM also construct a discriminant plane to identify the classes [26]. Figure 2.4 shows the hyperplane distinguishing two classes of data vectors with SVM.

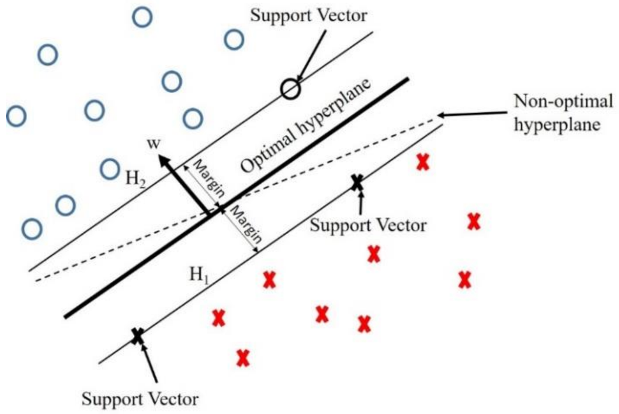


Figure 2.4. SVM classifier– uses the optimal hyperplane for generalization

2.2.2.2.2 k – Nearest Neighbour (k-NN)

In kNN Algorithm, the object is precisely classified by majority of its neighbours. The object is assigned to the most common class among its k nearest neighbours. For instance, the object is assigned to the class of single nearest neighbour, if k = 1. This algorithm is mathematical computational algorithm and is used for binary classification i.e., 0 & 1. In the feature space, input consists of the k closest training examples. k-NN classifier uses the technique of assigning the feature vector (unseen data point) to a dominant class among ‘k’ nearest neighbours. [27].

2.2.2.2.3 Naïve Bayes (NB)

NB Classifier is suited when the dimensionality is high & performs better with nominal data but not with numeric data. It considers contribution of each feature independently to the probability that the person has a heart disease. It is used for binary classification i.e., 0 & 1 since it is a mathematical computational algorithm. This algorithm is very stable as a small change in data set doesn’t make a remarkable change in the model.

2.2.2.2.4 Decision Tree (DT)

DT is a decision tool that uses a acyclic graph from training data to perform classification. Within tree structure, each non-leaf node is made responsible for testing a feature and the leaf node corresponds to a class label. The main advantage is the low computational complexity and disadvantage is that the constructed tree may become very complex at times.

2.2.2.2.5 Random Forest (RF)

RF consists of n number of decision trees, where each individual result is depicted as a separate tree. It is derived from the random decision of forest, proposed by Tin Ho in 1995 [28]. It selects attributes randomly to build decision trees with restricted fluctuations.

2.2.2.2.6 Logistic regression (LR)

If dependent variable is binary, LR is used for predictive analysis and dependent binary variable and one or more independent variables are relatable.

2.2.2.3 Heart disease prediction

Numerous studies have been performed for efficient diagnosis of heart ailments. Most authors applied various data mining tools for prediction and found out accuracies using different ML classifiers. According to Tougui, I et al [25], six data mining tools available nowadays for researchers are RapidMiner, Orange, Knime, Weka, Matlab, and Scikit-Learn. Moreover, six machine learning techniques; namely Logistic Regression, Support Vector Machine, K Nearest Neighbors, Artificial Neural Network, Naïve Bayes, and Random Forest, were also envisaged for classifying the presence of heart disease. Furthermore, three performance measures; i.e., sensitivity, specificity and accuracy, are being used to compare performance. In his case, Matlab was found to be finest performing tool because ANN model gave the highest accuracy of 85.86% in Matlab, and KNN gave the lowest accuracy of 63.64% in Knime's. In addition, few advanced methods for prognosis of cardiovascular disease were also identified by Georga, E. I. et al [29], showing that numerous intelligent data mining tools and ample ML algorithms are being used nowadays for modelling methods for calculating accuracy, precision and interpretability, incase of heart disease prediction.

H. Benjamin Fredrick David et al [30] evaluated performance of three data mining classification algorithms i.e Random Forest, Decision Tree and Naïve Bayes, with the help of heart disease dataset from UCI ML repository. Research found that RF algorithm performs best with 81% precision when compared to other algorithms for heart disease prediction, so that loss of lives may be avoided.

Mary, T. S. et al [31] focused on all significant features affecting heart disease and performed an analysis after varying set attributes to predict heart diseases. A gradual increase in prediction accuracies was found with the increase in number of attributes irrespective of classifiers. NB and RF algorithms performed better. Dataset-1 consisted of 7 attributes (age, sex, exercise induced angina, number of major vessels depression induced by exercise, slope of peak exercise, and type of defect). Dataset-2 had all attributes of Dataset 1 along with additional 3 more attributes (resting ECG, cholesterol and chest pain type). Dataset 3 had all 13 UCI hear disease dataset attributes. Holdout Validation was 70% split was used. In Dataset-1, NB algorithm outperformed with 78.02% accuracy. In the Dataset-2, NB algorithm outperformed with 82.75% accuracy and for Dataset-3, RF outperformed with an accuracy of 86.81%, as depicted in Figure 2.5.

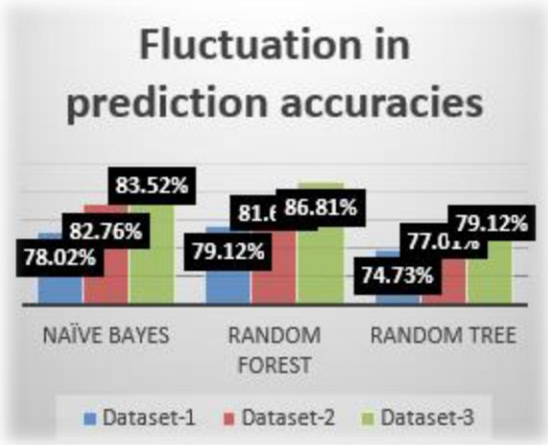


Figure 2.5. Fluctuations in prediction accuracies [31]

Similarly, Fahd Saleh Alotaibi et al [32], also used five different ML approaches (DT, LR, RF, NB, SVM) to understand data and predict the Heart Failure (HF), using the same UCI dataset. To avoid biasness, the random number generation method was used and data was enhanced by three times. The research also applied a feature selection approach and concluded that SVM gave 84.85% accuracy with standard dataset, and DT gave 93.19% accuracy with enhanced dataset, using 10-fold cross validation in RapidMiner environment.

Three different algorithms were proposed by Jyoti Soni et al [33] namely NB, KNN and DT, to predict heart disease. Among these, Naïve Bayes performed better with 10-fold cross validation. With Clustering as a pre-processing step and WEKA 3.6.0 as the data mining tool, dataset of 909 records with 13 different attributes, which were made categorical and inconsistencies resolved. DT outperformed other two data mining technique

Another research by S. Bashir et al, 2019 [34] focused on feature selection and ML algorithms where various heart disease datasets were used to demonstrate the accuracy. Different data mining techniques i.e., DT, LR, SVM, NB and RF were applied individually in Rapid miner on a UCI heart disease date set and compared the results. RF indicated 84.85 % accuracy in the results with 5 fold-cross validation.

In order to accurately predict presence of heart disease, Panda, D et al [35] also used seven MLalgorithms i.e. SVM, DT, NB, KNN, RF, Ensemble Classification (Extra Trees) and LR. The research pointed out that prediction accuracy is more when output model is trained with minimum no. of classes. Using original data set (Case 1) having classes 0, 1, 2, 3, and 4 in the predicted column, RF gave greater accuracy. However, after minimizing output classes to two (0 & 1) in the Case 2, Gaussian NB furnished greatest accuracy of 91.66%.

To improve performance of weak classifiers in predictive analysis of heart disease, few ensemble algorithms were selected by Latha, C. B. C. [36] e.g., bagging, boosting, voting, and stacking. Moreover, feature sets were named the FS1, FS2, FS3, FS4, FS5 and FS6. Ensemble techniques such as Boosting, Majority Voting, Bagging, and Stacking were used. The highest accuracy was obtained with set FS2 (containing 9 x features) & Majority vote with BN, NB, RF and MP i.e., 85.48%.

In order to ascertain prediction accuracy of heart disease from different attributes in heart-c.arff dataset, Weka software was used with naïve bayes and J84 classification algorithms by Gultepe, Y. et al [37]. Results were optimized through the use of the first ensemble algorithm, Bagging. However, optimization of J48 using Bagging increased the accuracy of classification reached to 81.31%.

To establish the classification accuracy of heart disease prediction, Khodke, H. E., et al [38] applied six different algorithms; namely, KNN, DT, LR, Gaussian NB, SVM and RF in Sci-Kit Learn tool. RF remained an effictive learning technique with the highest accuracy of 89.01%.

A study was proposed by Haq, A. U. et al [39] to develop an intelligent and hybrid ML predictive system for heart disease diagnosis. In this regard, seven popular ML algorithms (*K-NN, ANN, SVM, NB, DT, and RF*), three feature selection techniques (*minimalredundancy-maximal-relevance (mRMR), Least Absolute Shrinkage and Selection Operator(LASSO) and Relief*), and seven performance measures (*accuracy, specificity, sensitivity, Matthews' correlation coefficient, and execution time*), were utilized. Likewise, 14 attributes in the dataset were reduced to 6 attributes through these feature selection methods. Performance was validated with all features first and reduced set of features thereafter. Three feature selection(FS) algorithms, were also used to enumerate features that contributed the most in predicted value. LR showed the best accuracy of 89%, when selected by FS algorithm Relief and 10-fold cross-validation.

Ensemble classifiers were implemented by Emakhu, J., et al [40] for accurately predictng and early diagnosis of heart diseases using a subset of features. The dataset utilized was the Statlog heart disease. Performance measures, such as sensitivity, accuracy, and specificity, were used to evaluate performance. An accuracy of 87.04% through RF Ensemble Classifier was obtained.

To predict the prsence of the heart disease and to classify its risk level, Rajdhan.A, et al [41] proposed to implement four ML algorithms (NB, DT, LR and RF). Results confirmed that RF achieved highest accuracy of 90.16%.

In a comparative analysis of ML classifiers, Al-Mustafa, K. M. et al [42] proposed to classify heart disease dataset presence, using eight distinct ML algorithms (KNN, NB, DT J48, JRip, SVM, Adaboost, Stochastic Gradient Decent (SGD) and DT). Accuracy achieved by DT was 93.8537% employing all 13 attributes.

An Intelligent Heart Disease Prediction System (IHDPS) was proposed by Nidhi Bhatla et al [43]. 909 instances, obtained after enhancing Cleveland dataset, were divided into two equal datasets, i.e. training dataset (455 records) and testing dataset (454 records). Using 15 attributes (added 2 more attributes, i.e. obesity and smoking), three ML algorithms were discussed and NB proved to be the most effective with highest correct predictions (86.53%) for patients with heart disease, followed by NN (85.53%). On the contrary DT turned out to be most effective for predicting patients with no heart disease, i.e. (89%).

Debabrata Swain et al [44], applied binary classification in predicting heart disease to help medical practictioners in predicting heart disease risk, with 15 independent variables. Several binary classification algorithms were analyzed like LR, SVM, KNN, Gaussian NB, DT classifier and RF classifier. LR was found to be the best classification algorithm to classify the risk of heart disease with 88.29% accuracy.

For predicting the presence of heart disease, eight different supervised ML algorithms were used within WEKA environment by Spencer, R. et al [7], namely BayesNet, Logistic, Stochastic Gradient Descent (SGD), KNN (IBK in WEKA), Adaboost M1 with Decision Stump, Adaboost M1 with Logistic, Repeated Incremental Pruning to Produce Error Reduction (RIPPER or JRip in WEKA) and RF. It was found that BayesNet classifier, using Chi-squared feature selection technique, achieved the highest accuracy of 85.0%.

Taking lead from the available data mining techniques, various classifiers were studied by Wu, C. S. M. et al [19]. It was, therefore, determined that LR (86%) and NB (79%) gave better accuracy with high dimensional dataset and DT, RF perform better with small dimensional datasets. Moreover, RF was found better than Decision Tree Classifier due to itself being an optimized learning algorithm.

Insight of existing works related to machine learning and feature selection techniques, is presented in order to observe prediction of each algorithm. ANN, DT, Fuzzy Logic, KNN, NB and SVM are extensively used algorithms. Dire need for enhance classification accuracy in the early onset of heart disease [45], is still being felt.

Using WEKA as the data mining tool, Kanchan and Kishor et al [46] used NB, SVM, and DT classification algorithms to study supervised ML algorithms to predict heart disease. While Principal component analysis (PCA) was utilized for reduction in dimensionality, SVM outshined NB & DT.

Two different datamining tools (MATLAB© and WEKA©) using six distinct ML algorithms (Linear SVM, Quadratic SVM, Cubic SVM, Medium Gaussian SVM, Decision Tree and Ensemble Subspace Discriminant) are utilized by Ekız, S, et al [47] for classifying the heart disease. J48 & Subspace Discriminant performed better than all other algorithms . Linear SVM outperforms the cubic, quadratic and Radial Basis Kernel, but takes longer to process in both ML environments. Performance of DT is found to be better in WEKA and SVM performs better in Matlab. Overall, MATLAB turned out to be more flexible [47].

Four different Heart disease datasets namely Clevealand dataset, Hungarian dataset, V.A. dataset and Statlog dataset, consisting of 13 features each, were used with C4.5 and Fast DT for each data sets and computed the classification accuracy, in WEKA. El-Bialy et al [48] proposed five common features per dataset to form a New Pruned Dataset, which included only most commonly occurring features (ca, age, cp, thal) for prediction of heart disease. Results showed classification accuracy 78.06% is higher than using separate discrete datasets i.e 75.48% [48].

Alizadeh Sani, et al [49] designed a comprehensive web based database, encompassing 126 papers and 68 datasets; extracted from literature between 1992 and 2018 to aid the users. Data was collected to advance clinical diagnosis. In total, 140 ML or data mining methods applied to diagnose heart disease. This database is available in 3 formats:, Mysql script file (.sql), SqlServer script file (.sql) and comma-separated values (csv). A new dataset with minimum redundancy was interfaced with a web application to interact with this dataset.

With the aim of extracting the hidden information/patterns in medical data, a web-based ML application [50] was also developed to predict presence of heart disease, trained with UCI dataset, in order to find out the suitable ML technique. User just need to enter specific medical details to get the probability of presence of heart disease and subsequently result will be displayed. It was observed that Naïve Bayes had 60% accuracy, logistic regression gave 61.45% and SVM had 64.4%. Thus, SVM was selected being most efficient algorithm.

Nishanth Vaidya et al. 2020 [51]stressed on different AI algorithms for achieving highest accuracy in coronary illness. Data mining assisted them in choosing the treatment on the bass of experimenting on available data sources. In this regard, heart disease forecast method is proposed to identify illness precisely and it was found that accuracy of prediction turned out to be 94.60% with SVM when data set used with 75-25% holdout validation.

Researchers have used numerous blends of data manipulation, feature selection/extraction techniques, and classifiers for computing binary classification accuracy in precisely predicting presence of the heart disease with numerous traditional ML methods. Table 2.3 illustrates the summary of the classification accuracy published by researchers from time to time.

Table 2.3 Classfication accuracy for heart disease with Traditional ML Approaches.

Publication Year	**Patients/ Instances**	**Validation Type**	**Features Sel / No of Attributes**	**Classifier**	**Accuracy (%)**	**Reference**
2020	720		PCA Chi-squared 13	Bayes Net algorithm	**85.0**	[7]
2020	303 (UCI)	Holdout 80% -20%	13	Random Forest	**90.16**	[41]
2020	303 (UCI)	-	13	Random Forest	**89.01**	[38]
2020	1025 (Kaggle)	-	Classifier Subset Evaluator Up to 4 13	Decision Table	**93.86**	[42]
2020	303		13	Random Forest Ensemble Classifier	**87.04**	[40]
2019	303 (UCI)	-	13	SVM	**84.85**	[32]
	1013 (Enhanced UCI)	-	13	Decision Tree	**93.19**	
2019	303 (UCI)	-	13	Gaussian Naive Bayes	**91.66**	[35]
2019	4239 (Kaggle)	-	15	Logistic Regression	**88.29**	[44]
2019	303 (UCI)	-	13	Random Forest	**84.85**	[34]
2019	303 (UCI)		FS2 (9)	Majority vote with NB, BN, RF and MP	**85.48**	[36]
2019	303 (UCI)		13	Random Forest	**86.81**	[31]
			10	Naïve Bayes	**82.76**	
			7	Naïve Bayes	**78.02**	
2019	303 (UCI)		13	Bagging with J48	**81.51**	[37]
2018	303 (UCI)		Relief 6	Logistic Regression with C=100	**89.0**	[39]
2020	303		13	SVM	**94.60**	[51]

2012	909	15	Naïve Bayes through IHDPS	**86.53**	[43]

2.2.3 Bio-Inspired Data Mining Techniques

A bio-inspired data mining approach mimics biological nature and gets insight from the natural life to solve the real-life data extraction problems. These techniques are being widely used in fields of *intelligent computing* (virus or fault [52] detection, change detection [53]), *security* (anomaly [54] intrusion detection [55], damage detection [56], network monitoring), *engineering* (optimization [57], robotics [19]) and *artificial intelligence*. Even though bio-inspired ML techniques are being nurtured to subdue obvious limitations of traditional data mining methods [19], still; very petite work is seen in medical diagnosis based on predictive analysis.

2.2.3.1 Bio-inspired algorithms

Bio-inspired algorithms are the intelligent computation techniques, which are inspired by biological working mechanisms [19], and are broadly categorised in three categories: -

- Inspired from organism structure (e.g., Artificial Immune System (AIS), Neural Networks (ANN));
- Inspired from organism behaviors (e.g., Firefly and Ant Colony Algorithm);
- Inspired from evolution (e.g., Genetic Algorithm).

2.2.3.2 Artificial Neural Networks (ANN)

ANN is one of a bio-inspired ML algorithm, which is a combination of artificial neurons that produces non-linear decision boundaries. Multilayer perceptron (MLP) is the most widely used model with ANN. In addition to it researchers have applied several other bio-inspired classification techniques which include:

- Learning vector quantization (LVQ) NN [58]
- Fuzzy ARTMAP NN [59]
- RBF-NN [60]
- Bayesian logistic regression NN [61]
- Adaptive logic network (ALP) [62]

2.2.3.3 Heart disease prediction through bio-inspired techniques

As ML has opened new avenues in medical diagnoses, an analytical comparison for best available algorithm has been presented by Sharma and Rizvi et al [3]. Traditional Classification algorithms which were used are NB, SVM, DT, ANN, RF and LR. It was concluded that NB work well when data set is small, but not in case of bigger datasets. SVM proved fruitful when training time is quite less, and higher accuracy is desired. Nonetheless, ANN outperforms in accuracy; however, its execution time depends on number of layers.

Although bio-inspired classification, related to various diseases such as Diabetes [63], Parkinson Disease [64], Brain inabilities [65] is found in the literature, but petite attention has been drawn by the researchers towards heart disease prediction accuracy through bio-inspired ML techniques.

In order to compare predictive data mining approaches in medical diagnosis, 168 articles, were studied by Ghorbani, R. et al [66] related to heart disease between 1997 and

2018. Classification turned out to be the significant data mining techniques used by researchers in the literature. Moreover, Decision Tree and Bayesian Classifier method were found to be the most popular classification approaches, followed by a bio-inspired classification technique i.e. ANN based on biological nueral network. However, researcher have not determined the exact model used in ANN in most of the researches [66].

A.B. Akella & V. Kaushik [67] showed that six orthodox ML algorithms have achieved accuracies more than 80%, but the bio-inspired Neural Network proved to be the best algorithm to detect coronary artery disease (CAD), with accuracy of 93.03%.

Likewise, a hybrid NN, which incorporates a bio-inspired ANN and fuzzy neural network (FNN), was developed for classification of disease prognosis with UCI Cleveland heart disease dataset, by Kahramanli, et al [68] . Highest accuracy acquired, through 10-fold cross-validation, was 86.8%.

Similarly, heart disease prognosis was also done by Kemal Polat et al [69], using a bio-inspired fuzzy-Artificial immune recognition system (AIRS) classification algorithm along with a weighting scheme, which was based on KNN as a pre-processing step. This weighted dataset was then offered to AIRS with fuzzy mechanism. Accuracy achieved was 87.0 % [69].

In another novel approach for efficient prediction of heart disease, with Statlog (Heart) dataset, using a bio-inspired algorithm, was introduced by Sharma et al [70]. DT ML classification was utilized with four bio-inspired feature selection algorithms (binary firefly algorithm (FA), binary particle swarm optimization (BPSO), binary ant colony optimization

(ACO), and binary artificial bee colony (ABC)). Out of these, max accuracy of 90.09% was achieved using the BPSO feature selection technique.

Vijayashree, J [71].proposed an improved particle swarm optimization (PSO) with SVM using a novel fitness function as feature selection. Existing feature selection algos such as Chi-squared, Info gain, one attribute based, Consistency subset, Relief, CFS, filtered attribute, Gain ratio and PSO algorithm. Out of the several ML classifiers; such as NB, RF and MLP, PSO-SVM gave highest classification accuracy of 88.22% using Six selected features for prediction of heart disease.

Although very petite work has been done by researchers for prediction of heart disease through bio-inspired techniques, few researchers have used either the bio-inspired techniques as a feature selection/extraction method, or by using few classifiers in predicting presence of the heart disease. Table 2.4 shows summary of the classification accuracy published by researchers.

Table 2.4 Classfication accuracy for heart disease with Bio-inspired techniques.

Publication Year	**Patients/ Instances**	**Validation Type**	**Features Sel / No of Attributes**	**Classifier**	**Accuracy (%)**	**Reference**
2020	303 (UCI)	Holdout 70% -30%	13	Bio inspired NN	**93.03**	[67]
2020	303 (UCI)	Cross 10-Fold	13	Bio-inspired NN	**85.86**	[25]
2020	303 (UCI)	-	13	Deep learning NN using Talos optimization.	**90.78**	[72]
2020	270	-	Bio-inspired Binary particle swarm optimization (BPSO) 13	Decision Tree	**90.093**	[70]

2.2.4 Performance Measure of a Disease Prediction System

The performance of a disease prediction system is judged according to its disease classification accuracy, which is dependent upon the classifier. Classification accuracy is valid under two conditions:

- The classes are balanced, that is each class has same number data samples.
- The data is unbiased, that is each class exhibits the same performance.

If the data does not satisfy these conditions, then we can resort to confusion matrix which is considered to be an informative performance measure. Accuracy, sensitivity and specificity are used commonly in medical classification studies were used [68].

The performance of a classifier is computed on a pre-recorded data and hold out strategy is used. Hold out means that some part of data is kept aside and not used for training classifier and only used for testing and validation. However, some researchers use cross validation technique on training data which may result in overrated performance.

2.2.5 Bench Mark Research

Frequent efforts have been carried out by researchers to classify heart disease dataset with higher accuracies in the literature. The research is generally being carried out on standard widely used data set of UCI Cleveland Heart disease dataset, which is an open-source dataset available on UCI Data Repository website [24] for research. It has 303 patients/instances and 76 attributes. However, only 14 attributes are utilzed in all published literature. In this regard, work done by few researchers with highest accuracies of prediction, on same dataset is illustrated in Table 2.5.

Table 2.5 Binary Classification accuracy of bench mark approach

Benchmark Research	**A.B. Akella et al. (2020),** [67]	**Fahd S et al. (2019),** [32],	**Al-Mustafa, et al. (2020),** [42]	**Nishanth Vaidya et al. 2020** [51]
Dataset	UCI Cleveland HD	Enhanced UCI Cleveland HD	Kaggle Enhanced HD Combination	UCI.
Patients/Instances	297	1013	1025	303
Pre-processing	6 discarded			-
Feature Selection Technique			Classifier Subset Evaluator	-
Features/Attributes	13	13	13	13
Validation Type	70-30 % Split	10-fold cross	10-fold cross	Holdout 75-25%
Technique used	Artificial NN	Decision Tree	Decision Table	SVM
ML Platform	R	RapidMiner	WEKA	-
Accuracy in %	**93.03**	**93.19**	**93.8537**	**94.60**

2.2.6 Gap Identification

Background and Literature review divulges that enhancement of classification accuracy for heart disease prediction, by means of ML, has always been a challenge for researchers. It has been found that Original UCI dataset is either enlarged/enhanced or dimensionality reduction techniques used to reduce no of features to get the desired or better accuracy results. Similarly, ample approaches have been used for achieving higher classification accuracy which include techniques of pre-processing, feature selection/dimensionality reduction is adapted through Relief, and BPSO before application of classification techniques. Nowadays, bio-inspired methods are being used as a feature selection/reduction technique, but at a very limited scale. Keeping the dataset intact i.e., without enhancing the number of patents/instances, incorporating all the features in the dataset and most of all using bio-inspired classification and optimization at the same time, very few researchers have achieved the desired results. As a summary, for achieving the higher classification accuracy, typical process of training and testing phases of classifier for heart disease diagnosis is illustrated in figure 2.6

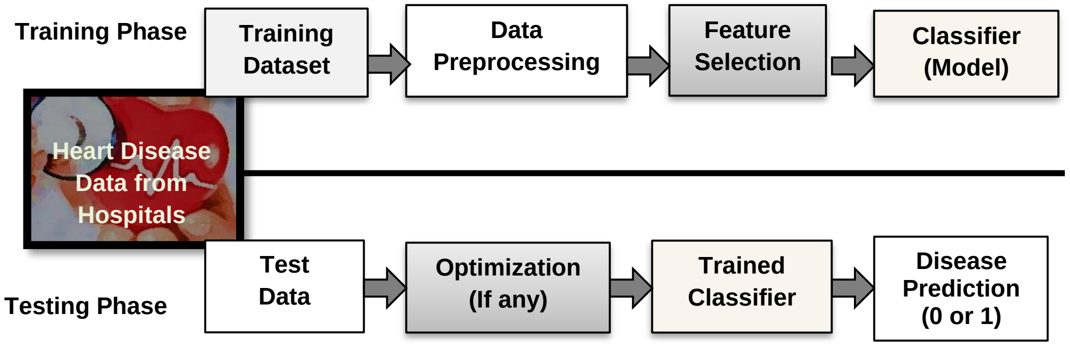


Figure 2.6. A typical classification process in prediction of heart disease

In the light of literature review, Heart Disease Prediction through classification techniques have focused on development of an efficient algorithm, which is desired to be robust and predict more accurately even for small training samples data sets. Since bio-inspired techniques are not widely used and ample traditional ML algorithms are available for study, the challenge is to develop a robust bio-inspired classification system for the precise prognosis of the heart disease presence. Moreover, the comparison of classifiers can only be done if it is used with, same dataset without enhancing dataset or reducing features, by some other researcher. For this purpose, standard open-source datasets are available which are used by researchers for comparison of effectiveness of algorithm with other techniques.

“Our research aims at to develop an efficient bio-inspired method (algorithm) for improvement in classification accuracy for prediction of heart disease.”

To achieve the aim, we have used open-source Cleveland Heart Disease data set from UCI Data Repository in this research to test our proposed methodology.

Chapter 3

THE IMMUNE SYSTEM

This chapter presents a little background of bio-inspired acuities, which are being dispensed in this research. It not only includes the introduction to Biological Human Immune System (BHIS) along with Artificial Immune System and few bio-inspired algorithms currently being used.

3.1 Biological Human Immune System (BHIS)

Biological Human Immune System (BHIS) is a defence system hosted inside our body which comprises of many biological cells, structures, molecules and processes that protect us against disease. BHIS in itself can be considered as an efficacious classification system which differentiates “good” i.e., S (body’s’ own organisms) and “not good” i.e., NS (intruders) [55]. It is imperative for BHIS to not only detect pathogens, or a foreign microorganism that can cause disease, but also differentiate between them and own body healthy tissues. Pathogen have the capability to change the configuration to evade detection by immune system, however, at the same time our body, possessing multiple defence mechanisms, can grow and recognize such pathogens. The host defence mechanism consists of *non-specific innate* (native/natural) immunity and *adaptive* (acquired/ specific) immunity.

The first line of defence in immune system is provided by innate immunity comprising of epithelial barriers and specialized natural antibiotics to stop microbes, bacteria, and viruses entering into our body. It has the ability to identify the bacteria (intruders) and neutralize them through pattern recognition receptor (PRR) sets. Innate immunity possesses capability of discriminating between S and NS molecules and stimulate a signal to antigen presenting cells (APCs), which leads to activation of T cells. Henceforth, the adaptive immune response starts. In case microbes succeed in breaching epithelia, phagocytes (natural killer cells) come in action and engulf them.

Adaptive immunity is divided into antibody humoral immunity (with B lymphocytes) and cell mediated immunity (with T lymphocytes [73]). Both B and T-lymphocytes are part of adaptive immune system, which recognizes the pathogens and make a bond with them for further processing to kill them by immune system. This recognition is done through antigen receptors (antibodies) which are present on the surface of lymphocytes. The receptors over these lymphocytes are produced and trained by our immune system to match with antigen and recognize it. The recognized antigen is then further processed for elimination and destruction by our immune system. In this way, the body responds to any microbe entering into the body, even if that microbe has never been seen by the immune system.

Backup immunity or passive immunity can be developed artificially in body. This differs from innate and adaptive immunity as it is acquired artificially through vaccine and can last for short or in some cases for long (lifetime) duration e.g., vaccination given to small babies for lifelong protection or Plasma transfusion to critically ill COVID-19 patients until effective medications is available [74]. It is also useful for immunosuppression as in case of an organ transplant; so that the transplanted organ can be accepted by the body. The multilayer defence system of BHIS is shown in figure 3.1.

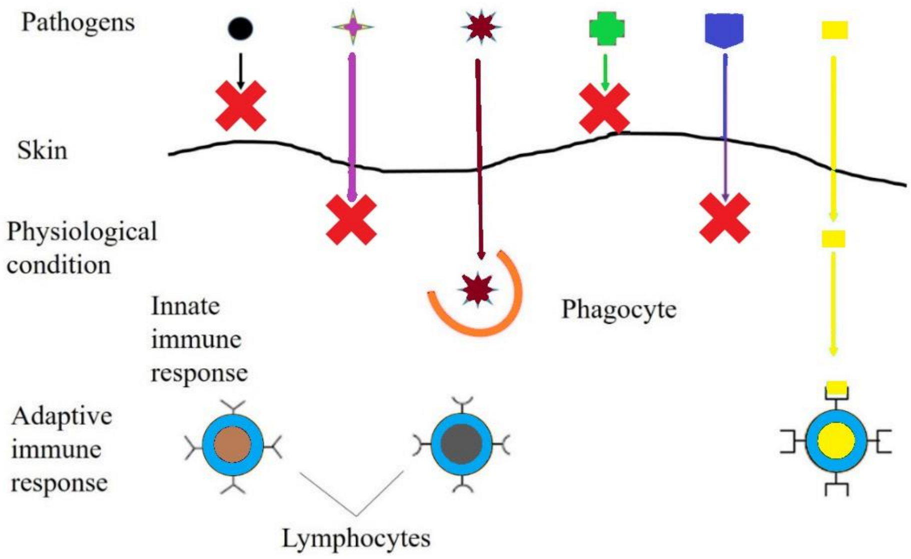


Figure 3.1. BHIS multilayer structure
Three layers of defense against antigens 1) Skin, 2) Innate response and 3) Adaptive response **[75]**

3.1.1 Physiology of Immune System

As BHIS has numerous types of immune cells, but our focus will be on lymphocytes only, which are the type of white blood cells, working against infections and protecting human body from intruder cells (bacteria and viruses). Bone marrow and thymus are the two organs of human body that are the source of generation and development of lymphocytes. B cells are developed and matured in bone marrow and T cells are migrated to thymus and matured there. Both type of cells has receptor molecules on their surface which can recognize antigens due to their specific patterns. Figure 3.2 illustrates the B cell and T cell receptors embedded on the surface of cells and in detached form for more details. T cell receptor is called TCR and in the same way, B cell receptor is called BCR or antibody (AB).

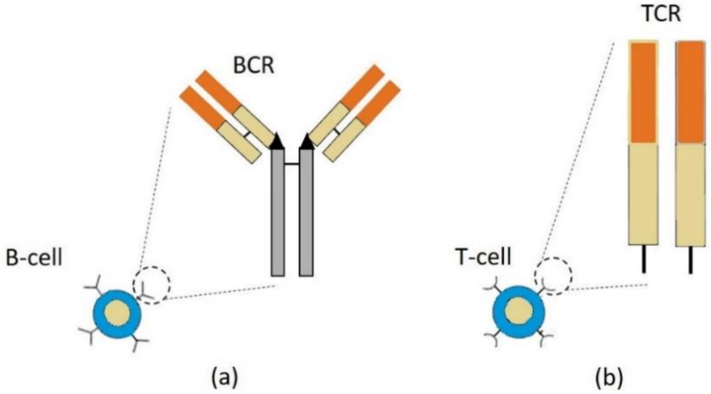


Figure 3.2. Lymphocytes with surface molecules
(a) B-cell and BCR, (b) T-cell and TCR [76]

Both cells have receptors molecules at their surface and their function is to recognize (bind) with the specific antigen. Antibody molecules and TCR are generated through the process of DNA re-arrangement in bone marrow [77]. Basic difference between B and T cells is that antibody molecules can recognize and bind with antigen even if these are free from B-cells, however, in T-cell the TCR can recognize the antigen and triggers an immune response. Moreover, the maturation place for B-cells is bone marrow and for T-cells it is thymus.

BHIS can recognize any antigen due to its specific pattern (shape) with the help of lymphocytes (B cells and T cells). Figure 3.3 explains the mechanism of pattern recognition by a BCR and TCR. Antibodies can recognize antigens even when detached from B cells and floating in blood stream. TCR recognizes those antigens, which are further presented to T cells by our body molecules called major histocompatibility complex (MHC).

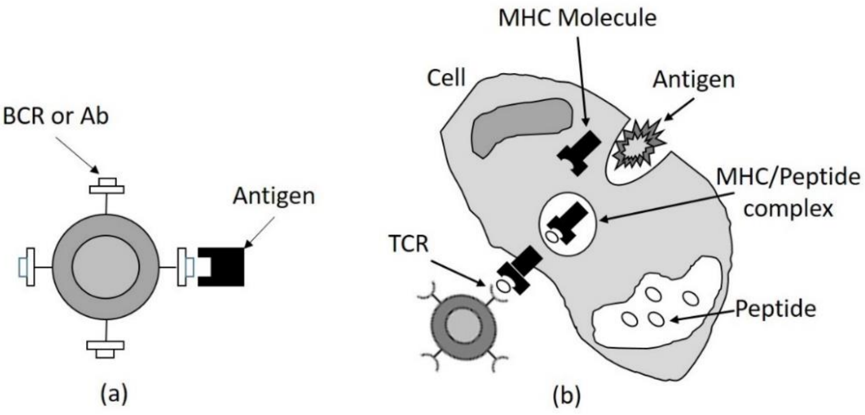


Figure 3.3.BHIS - pattern recognition
(a) Pattern recognition by BCR (b) Pattern recognition by TCR [77]

3.1.2 Thymic Negative Selection – self (S) and non-self (NS) discrimination

T-cells are matured in thymus, the process of thymic negative selection of T-cells is carried out during maturation. During this process, S-cells of body (S-antigens) are presented to T-cells, if any of the T cell recognizes and binds with S-antigen it is eliminated and destroyed in thymus [78], [79]. T-cells are generated in bone marrow and migrated to thymus, located in upper region of chest as shown in figure 3.4.

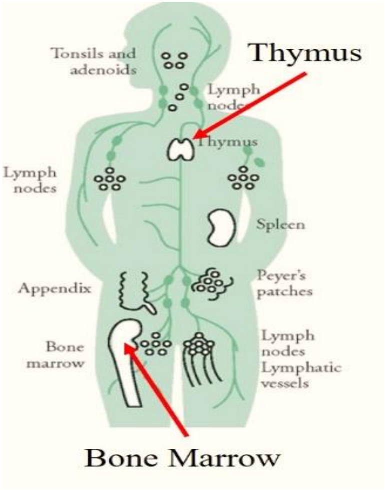


Figure 3.4. Bone marrow and Thymus [77]

Self antigens (in thymus) are presented to immature and naïve T-cells, which undergo a negative selection. If naïve T-cell identifies any of the presented S-antigen, it is eliminated from the population of T-cells. The naïve T-cell that does not identify and binds with any of the presented S-antigen, thus, becomes an immunocompetent or mature T-cells. These matured T-cells after the negative selection methods in thymus are allowed to go in the blood stream and patrol the body in search of NS antigens (which are the intruder cells) thereby performing an immune response. Figure 3.5 describes the process thymic negative selection and maturation process of the T-cells.

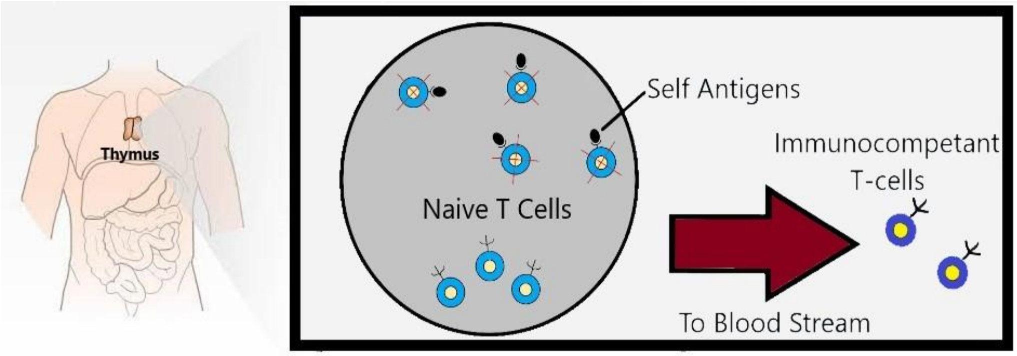


Figure 3.5. Thymic negative selection and maturation process of T-cells

The recognition followed by binding of lymphocytes and mutation cells with antigens can be a measure of affinity maturation, which is referred to as degree of binding between the cells and antigen. This means that a higher affinity measure will depict a stronger bond between T-cells and antigens and results in a better recognition and immune response. As this response is through mutation followed by selection, it is called an adaptive response, which guarantees successive encounters with certain type of antigens. The higher the affinity of a receptor cell with antigen, lower will be its mutation rate during maturation of T-cells and vice versa.

Inverse to mutation, proliferation rate (increase in number) of a cell is directly proportional to its affinity with the antigen. When a non-self (NS) antigen invades the body and is recognized as antigen by immune cells, these undergo clonal selection and affinity maturation. Off-spring cells are increased in numbers and it is directly proportional to the affinity with antigen.

3.2 Artificial Immune System (AIS)

Artificial immune system (AIS) is composed of intelligent methodologies, inspired by BHIS, for finding solutions to real-life problems. It not only mimics adaptive capability of BHIS, but also follows its ability to recognize and memorize complex patterns [75]. Uniqueness, pathogen recognition, memory and reinforcement learning are those characteristics of BHIS, which are quite suitable for the development of systems which are dynamic, distributed, self-organized and autonomous [80].

If each element of a vector represents a feature in a data set and located in a feature or “shape space”. Vectors are normalized between [0, 1] in order to visualize the data in shape space, if plotted into a defined dimensional space. Artihematically, a molecule (mo) of AB or AG are being represented by a set of coordinates $= [mo_1, mo_2, \ldots \ldots \ldots \ldots mo_{n]}\ where\ mo \in S^n$. Affinity between AB and AG is the measure of distance between the two vectors. Euclidian distance, is being used here to calculate it. If the coordinates of an AB are given be $\mathrm{AB} = [\mathrm{AB}_1, \mathrm{AB}_2, \ldots \ldots \ldots \ldots \mathrm{AB}_{L]}$ and AG are given by $\mathrm{AG} = [\mathrm{AG}_1,\ \mathrm{AG}_2, \ldots \ldots \ldots \ldots \mathrm{AG}_{L]}$, then the distance E_Dist can be represented by the equation (1)

$$E\_Dist = \sqrt{\sum_{i=1}^{n} (\mathrm{AB}_i - \mathrm{AG}_i)^2} \quad (1)$$

3.2.1 Negative Selection Artificial Immune System (NSAIS)

Nowadays, AIS based ML models ashore on Danger Theory, Clonal selection, Negative selection, Immune network and Adaptive Immunity. In this study, we restricted ourself to thymic negative selection only in BHIS. Thymic negative selection was originally documented for detection of anomaly with the help of detectors, in order to distinguish between normal and abnormal data [53], utilizing the self-nonself (S-NS) discrimination notion. These detectors are the randomly generated normalized real valued vectors in the NS-space, with dimensions exactly equal to the dimension of training sample (S-sample). This concept of NSAIS is shown in figure 3.6.

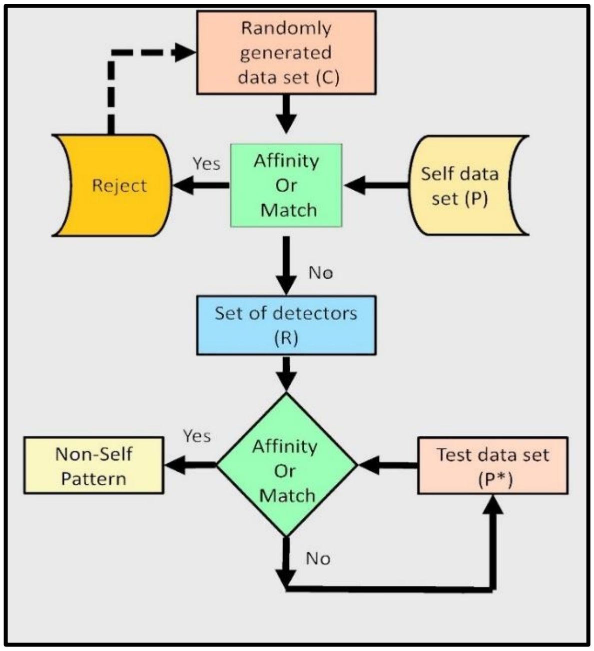


Figure 3.6. Negative Selection of Detectors [65]

In the generation stage, random detectors are generated and checked if they recognize S-samples or otherwise. The affinity between each S-sample and all randomly generated candidate detectors is checked through the equation (1). The detector which matches with Self-sample is discarded. A new detector is randomly generated. If detector does not match with Self-sample, it is immediately contemplated as a nonself detector and added to detector set.

In detection stage, affinity of selected NS-detectors set through the test data set is computed. Negative pattern detection algorithm uses the matching rule. Euclidian distance is used for binary representation of real valued data. If the distance between the sample and detector is less than a certain threshold (σ), matching occurs. For a perfect matching, the two patterns are equal however, in real valued application partial matching concept is used as If the Euclidean distance, $E\_Dist(AB, AG) = 0$,, the two patterns are equal and perfect matching takes place. In case of $D(Ab, Ag) < \sigma$, partial matching concept is used. The threshold used by Euclidian distance is radius of the detector.

The main purpose of the detector generation is to cover the non-self space effectively, without any voids, however, limiting the number of detectors and radius of S-samples may not allow the effective coverage. Figure 3.7 shows Self region along with Nonself region with detectors. Yellow portion divulges S region and blue discs represent detectors covering the NS-space. A detectors set is required for effective coverage of complete NS-space.

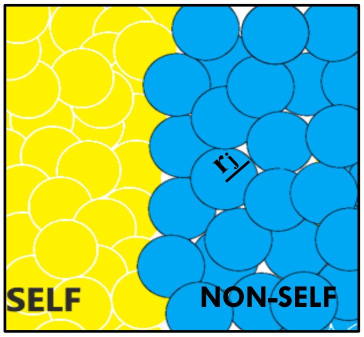


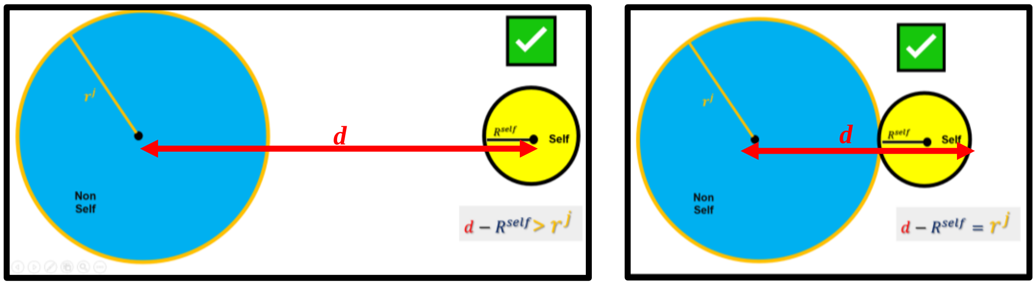
Figure 3.7. Self-Nonself region in Vector Space

Detector generation in NSAIS initiates with random candidate detectors, which are, thereafter, matured through an iterative process. In random generation, centre of each detector is chosen randomly and its radius is a variable parameter, which represents the size of the detector (can also be termed as influence region of detector). Influence region of each candidate detector is calculated in each iteration and checked that it should not fall within Self-region.

For a detector to be valid, two conditions must be fulfilled, i.e detector discriminates between S-samples either by remaining away from S samples or near boundary, as shown in equation (2) and figure 3.8.

$$d - R^{self} \geq r^{j}, \tag{2}$$

$$where\ \ 'd' is\ the\ distance\ between\ centre\ of\ the\ self\ sample\ and\ the\ centre\ of\ the\ random\ derector$$

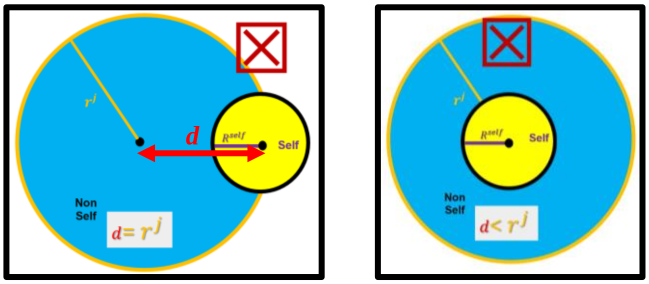


(a) (b)

Figure 3.8. Matured Detectors.
(a) Self Sample at a distance d from Non self detector (b) Nearest Self Sample from Non self detector

If this condition is not fulfilled and the detectors are either coinciding with S-samples or are within the S-region, the detectors will be discarded as shown in equation (3) and depicte in figure 3.9.

$$d \leq r^j \tag{3}$$



(a) (b)

Figure 3.9. Discarded Detectors

(a) On the boundary (b) Within the non self space

3.3 Genetic Algorithm (GA)

In 1960, Genetic Algorithm (GA) was first proposed by John Holland in 1960s [81], stimulated by Darwin’s theory of natural selection [57]. The entities with highest fitness, are selected for further reproduction to produce better off-spring for the upcoming generations. The higher the fitness value, the higher the ability of the computational chromosome to solve the real-world problems. This notion allows GA to find optimal solutions, even if the dataset has discontinuities. GA's fitness function evaluates and selects the best chromosomes for a upcoming population. These chromosomes are represented in binary strings, with two possible alleles for each chromosome's locus: 0 and 1. Similarities between biological chromosomes in natural evolution and computational chromosome in GA are as illustrated in Figure 3.10.

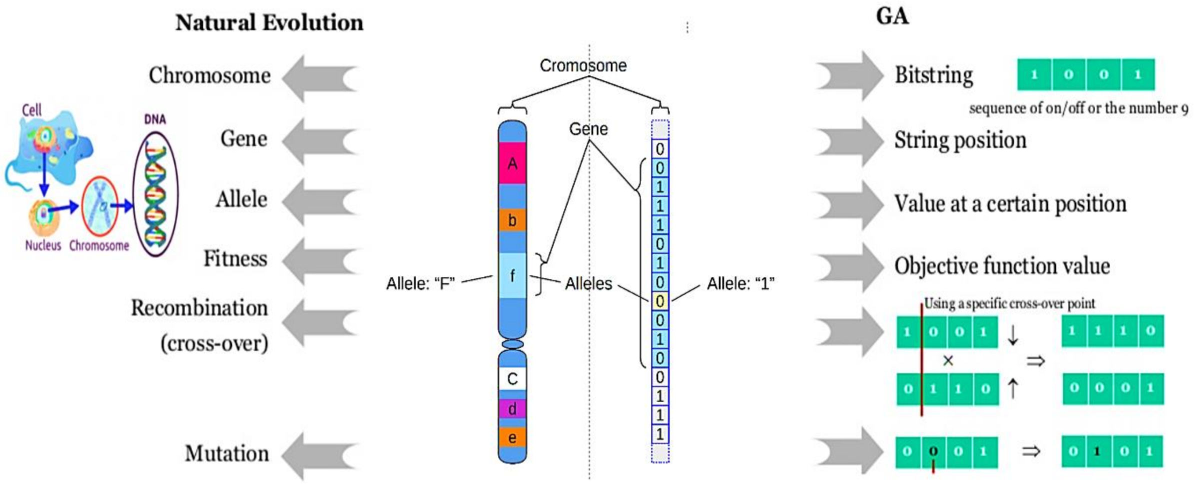


Figure 3.10. Biological Chromosome vs Computational chromosome [82].

3.4 Applications of NSAIS & GA

Over the past few decades NSAIS has been explored and used by the researchers in the field of artificial intelligence. The artificial recognition system (AIRS) has been used and exploited as binary classification method based on Biological Human Immune System (BHIS) response against pathogens [83]. The way BHIS distinguishes between the body's own cells or organism i.e., good and the outsider/foreign cells i.e., not good, NSAIS differentiates S and NS in the same way [55]. Likewise, anomaly/intrusion detection [54] [55] can be defined as detection of abnormal examples from normal examples. Computer security and virus detection could be classified as anomaly detection [84]. Application of AIS has also been seen in fields of *intelligent computing* (virus or fault [52] detection, change detection [53]), *security* (anomaly [54] intrusion detection [55], damage detection [56], network monitoring), *engineering* (optimization [57], robotics [19] [85] and isolation on motor bearings [86]).

In the same way, Genetic Algorithm (GA) is being mostly used in healthcare as feature optimizing technique for efficient classification disease data set [64]. Numerous researches have used GA as a feature optimizing technique or as a hybrid algorithm with fuzzy systems or NN. Being an optimization technique, it is mostly used for exploring possible optimal solutions.

Chapter 4

PROPOSED METHODOLOGY

This thesis primarily focuses on a bio-inspired classification model based on negative selection framework, to classify the disease data for accurate prediction of the heart disease. The ultimate goal is to extract reliable information from available data, using information to train a bio-inspired classification system and finally make decision (assign distinct class) to the test data, in order to correctly classify presence of heart disease. In addition to it, heart disease prediction through multiple traditional machine-learning classification models were also implemented results in two ML platforms i.e., MATLAB and WEKA, and then results were compared. The step-wise development of proposed classification algorithm is discussed in this chapter.

4.1 An Overview of Proposed Methodology

Since the heart disease is primary cause of death globally and colossal healthcare data is depleted in hospitals, achieving best classification accuracy in prediction models has always been a challenge for the researcher. In this research, study of existing traditional ML methods (Mode-1), along with a novel bio-inspired technique based on negative selection artificial immune system (NSAIS); capable of classifying classify, optimize detectors and finally predict presence or absence of the heart (Mode-2), is proposed.

4.1.1 Description of Data Set

Cleveland UCI Heart Disease Data repository [24] has been used in this research explained in the Section 2.2.1. It contained 14 attributes and 303 patients. Out of which, 54.1%

(164) did not have heart disease, while 45.9% (139) had heart disease. Binary classification of this data sets shows absence (‘0’) and presence (‘1’) of disease.

4.1.2 Data Pre-processing

In this research, widely used UCI Cleveland Heart Disease Dataset is utilized. Original dataset had 5 Classes (0, 1, 2, 3, 4) but for this study binary classification (0,1) is used. Out of 303 patients, six missing values were found in the data set, which were pruned to give final dataset of 297 instances, which were normalized for further dispensation.

4.1.3 Validation Techniques

4.1.3.1 Holdout Validation

Hold-out is to split the dataset into a *train* set and a *test* set. During the learning and testing phase, 70-30 percentage split was used & results were logged.

4.1.3.2 Cross Validation

Cross validation divides the data into k equal subsets by giving chance to each subset in training and testing phase. Hence, the testing data selection is different every time and compiles the average results at the end. 10 folds Cross Validation was used for validating the data.

4.2 Mode 1-Traditional ML Classification

In this mode, we not only applied five traditional ML algorithms (KNN, NB, RF, LR and SVM), but also compared their performance in two different ML platforms with the same UCI dataset, and subsequently compared the results for the highest classification accuracy in prediction of heart disease. At first the model is trained on training dataset, and then the test set evaluates how well that model performs, when exposed to unseen data. This research aims at achieving highest classification accuracy for the prediction of heart disease. Henceforward, Mode-1 can be divided into three stages as under: -

- Stage 1). Dimensionality Reduction through Principal Component Analysis (PCA)
- Stage 2). Training Classification Model with Traditional ML Techniques
- Stage 3). Presenting test data to Trained traditional ML Model.

4.2.1 Dimensionality Reduction through Principal Component Analysis (PCA)

In order to achieve a high accuracy for correctly predicting presence of heart disease through the symptoms (features or attributes) using different algorithms, various parameters are tested for different values. The attribute selection was done through PCA which reduced dimensionality up to 1 attribute (PCA enabled) from all the 13 available attributes (PCA disabled). PCA reduces the dimension of large data sets. PCA is such a tool, which reduces higher dimensional dataset to a minor dimensionality to discover any hidden insights. During model learning and testing phase the dimensionality reduction up to 5 attributes was used. Results using PCA disabled were also recorded, which meant to use all the 13 available attributes.

4.2.2 Training Classification Model with Traditional ML Techniques

After preprocessing, holdout validation of 70-30 percentage split was applied. Thereafter, the performance of all five algorithms was studied, one by one, by first training them with train data set and then evaluating their accuracy based on test data set in Matlab and Weka independently.

4.2.3 Heart Disease Predicion Through Trained traditional ML Model

Once the model is trained with the available test data (70%), the test data is given as an input to the trained *traditional ML Classification model* to correctly classify the presence or absence of the heart disease. In this regard, workflow of the traditional ML methodology is illustrated in Figure 4.1.

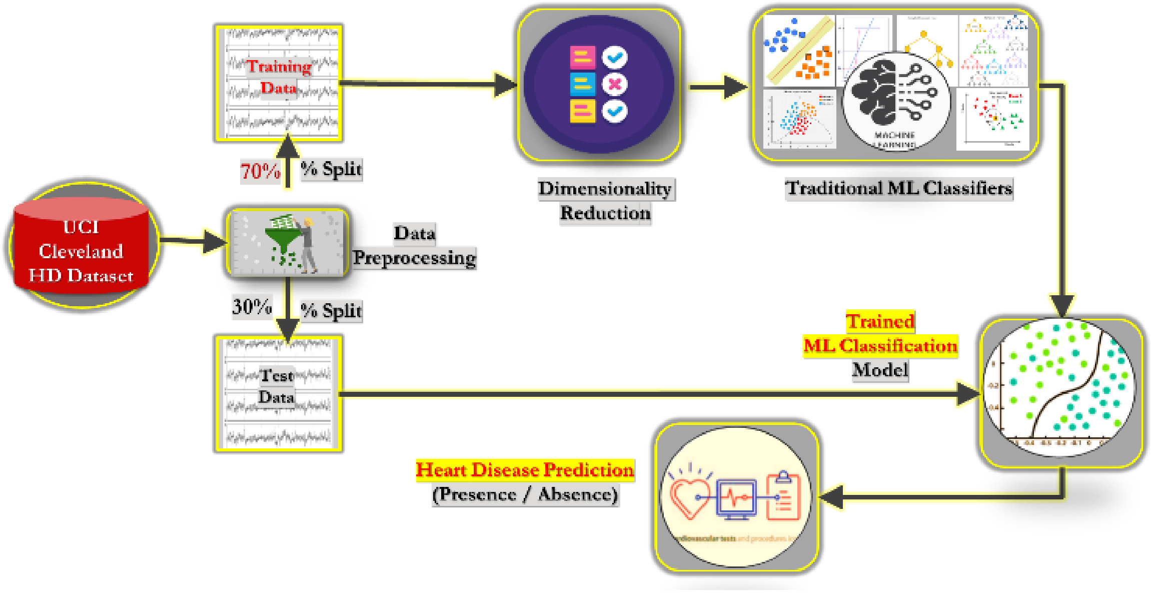


Figure 4.1. Traditional ML Methodology

4.3 Mode 2 – Proposed Bio-inspired Classification (BIC) based on NSAIS

Our proposed Bio-inspired Classification based on Negative Selection Artificial Immune System (BIC based on NSAIS) replicates S-NS discrimination notion, in biological thymic negative selection principle. As T Cells are nurtured and matured in the thymus, similarly, the random candidate detectors are also generated in a defined vector space, to cover the non-self-region. Likewise, the way Naive T Cells are matched with the self-antigens, our real valued non-self-detectors are also assessed on the basis of the affinity with Self Samples.

Self-matching detectors are eliminated/discarded. On the other hand, the detectors, which do not recognize the self-samples are selected and these *matured detectors* are stored in a detector sets; just like the matured/immunocompetent T cells. General resemblance between the Bio-inspired Thymic Negative Selection in BHIS and our proposed NSAIS is exemplified in figure 4.2.

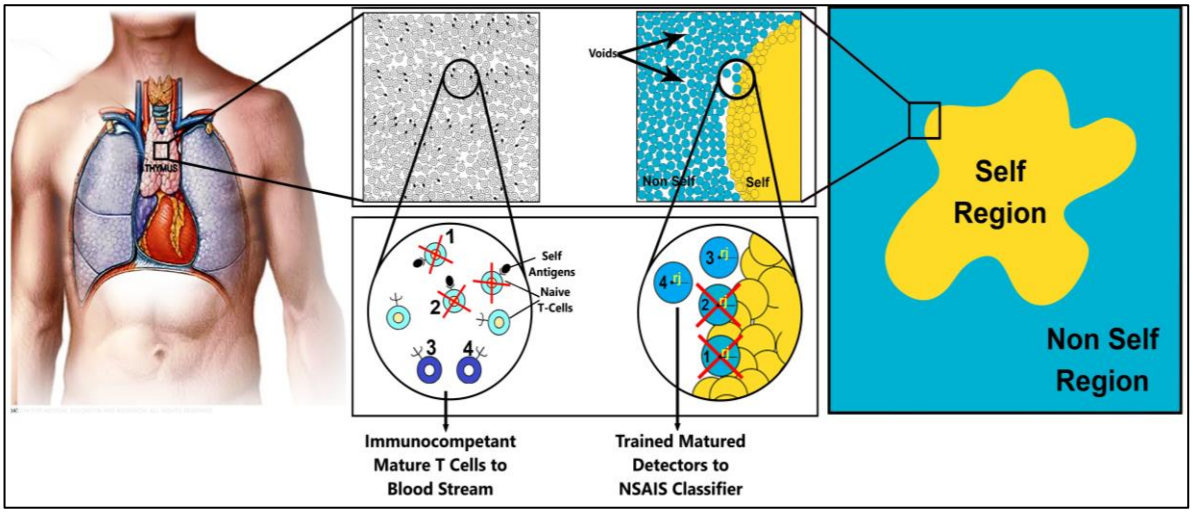

THYMUS
Voids
Non Self
Self
1
2
3
4
Self Antigens
Naive T-Cells
4
3
2
1
Self Region
Non Self Region
Immunocompetant Mature T Cells to Blood Stream
Trained Matured Detectors to NSAIS Classifier

Figure 4.2. Impersonation between Thymic Negative Selection BHIS and NSAIS

Our proposed methodology (Mode-2 BIC based on NSAIS), can be divided into five stages listed as under: -

- Stage 1). Pre-processing of Cleveland Heart Disease Dataset.
- Stage 2). Holdout Validation with Percentage Split 70%-30%
- Stage 3). Random detector generation through NSAIS.
- Stage 4). Optimization of detectors sets using Genetic Algorithm (GA).
- Stage 5). BIC based on NSAIS
- Stage 6). Presenting test data to Trained BIC based NSAIS Model

4.3.1 Random Detector Generation through NSAIS

The concept of Bio-inspired negative selection is used for generation of NS detector set for effective coverage of NS-space. The total number of detectors play a pivotal role of adequately covering the NS-space. A reduced number of detectors may result in inadequate coverage of NS-space while increased number of detectors may result in overlapping of detectors and detection accuracy is reduced due to false negatives. The detector set exhibiting the highest (best) detection accuracy is selected and used for classification of heart disease.

As we proposed random detector generation in the normalized feature space, self-matching detectors are eliminated using negative selection algorithm (NSA). At first, the detectors are generated to cover the compete non-self region in a defined vector space, as shown in the Figure 4.3.

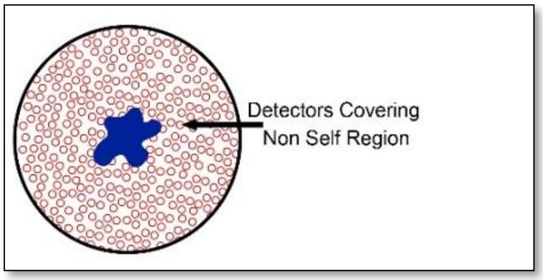


Figure 4.3. Detectors with void in non-self region

However, there is still a chance of presence of redundant detectors within the non-self detectors, hence the overlapping detectors are also removed, so that the non-self space can be covered by minimum number of detectors. Thus, the detectors are trained using the S-sample (training sample), and the best optimized detector set are saved effectively covering non-self space for the particular class, while avoiding voids and mutual overlap as shown in Figure 4.4.

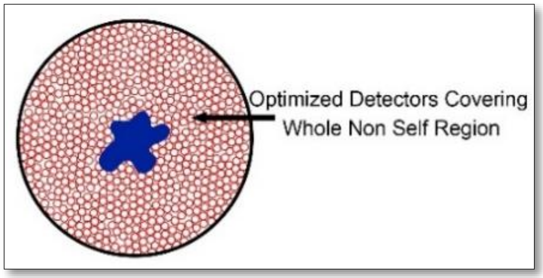


Figure 4.4. Optimized Detectors in NS region

4.3.1.1 Detector Generation Algorithm

Detector generation algorithm is shown in figure 4.5 [87]:

```
Detector Set ( $\omega$ , $Dn, R^{Self}$)

$\omega$: $set\ of\ self\ samples$

$Dn$: $number\ of\ detectors$

$R^{Self}$: $radius\ of\ self\ sample$

1: $D = \emptyset$

2: $Repeat$

3:    $x = random\ detectors\ generated\ [1,0]^n$

4:    $Repeat\ for\ every\ \omega_i in\ \omega = [\omega_i, i = 1,2, \dots \dots \dots]$

5:          $d = Euclidean\ distance\ between\ \Omega_i\ and\ x$

6:          $if\ d \leq R^{Self}, go\ to\ 2$

7:    $D = D \cup [x]$

8: $Until\ |D| = \ Dn$

8: $Return\ D$
```

Figure 4.5. Detector generation for NSAIS

4.3.2 Optimization of detectors sets using Genetic Algorithm (GA).

In this thesis, GA has been proposed to get an optimal solution by generating optimized number of detectors to cover the NS-space. The optimization of detector set comprises of following steps:

- Selecting Fittest detectors
- Roulette wheel selection
- Multi point crossover
- Bit-wise(bit-flip) mutation

GA optimized negative selection detector optimization is shown in figure 4.6.

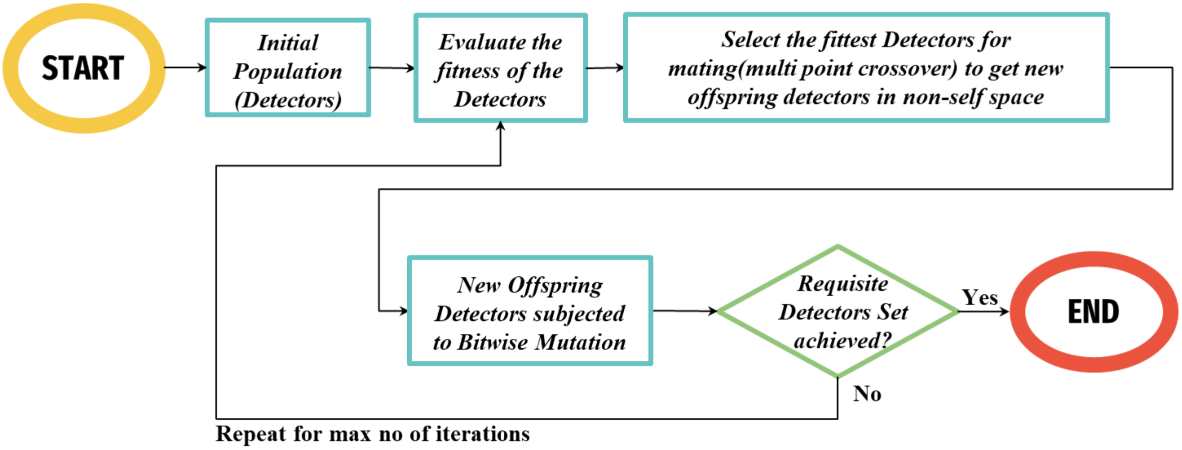

START
Initial Population (Detectors)
Evaluate the fitness of the Detectors
Select the fittest Detectors for mating(multi point crossover) to get new offspring detectors in non-self space
New Offspring Detectors subjected to Bitwise Mutation
Requisite Detectors Set achieved?
Yes
END
No
Repeat for max no of iterations

Figure 4.6. GA optimized negative selection detector optimization

The selection of the best detector set is based on the accuracy. Step-by-step optimization of detectors using GA is described as under:-

- **Step 1:** Maximum coverage of non-self space by detectors and determining optimal number of detectors.

- **Step 2:** Determine centre location of detectors and influence region (r^j).
- **Step 3:** Optimization criteria including
  - Minimization of detector overlap to reduce number of detectors
  - Maximization of detector diversity to reduce redundancy
- **Step 4:** Use of optimized/matured detector set for classification using NSAIS.

4.3.2.1 Optimization of detector radius

Let there are "N" number of k dimensional S-samples. NSAIS aims at to generate k dimensional detectors which cover maximum NS-space with fewer number of detectors. For Real Valued NSAIS, the i^{th} self sample ω (training sample) and j^{th} detector d can be defined by the set of training data (S-samples) is defined as a vector $\omega^i = [\omega_1^i, \omega_2^i, \omega_3^i, \dots\dots\dots\dots \omega_k^i]$ with Self-radius $R^k = R^{self}$ and set of detectors is defined as a vector $d^j = [d_1^j, d_2^j, d_3^j, \dots\dots\dots\dots\ d_k^j]$ with detector radius r^j. A two-dimensional representation of Self samples and detectors was also shown in figure 3.6

Consider that the location of self centre of i^{th} S-sample is given by $Self^i$ and the location of centre of j^{th} detector is given by d^j. Also consider that $Self^j$ is the nearest S-sample from d^j. The detector location is to be optimized so that it can cover the maximum NS-space and at the same time should have minimum overlap with other detectors. The detector's distance from its nearest S-sample is given by equation (4)

$$\left|Self^i - d^j\right| = \underset{i}{\overset{N}{min}}\left|Self^i - d^j\right| \qquad (4)$$

Detector’s maximum radius can be given by equation (5)

$$r^j = \left|Self^i - d^j\right| - R^{self} \qquad (5)$$

Where,

$$R^q = \left|Self^i - d^j\right| \; or \sqrt{\sum_{k=1}^{k}(\omega_k^i - d_k^j)^2}$$

Now for any detector d^j with radius r^j, if there is any S-sample $Self^i$ with radius R^{self} the validity of detector can be given as

$$if \begin{cases} R^q \leq R^{self} \;(\,invalid\ detector, overlaps\ self - sample) \\ \qquad valid\ detector, otherwise \end{cases}$$

Therefore, the maximum possible radius a valid detector can have will depend upon its distance from nearest S-sample. This gives rise to the fact that the detector located in close vicinity of S sample will have small radii while the detectors located at far distance can have large radii. Selection of a detector can be given by equation (6)

$$r^j = \underset{d^j}{max} \; |R^q| - R^{self} \qquad (6)$$

The basic objective is to maximize the radius of detector, but it depends upon the radius of nearest S-sample and nearest S-distance. The radius of S-sample is kept fixed. If the radius is kept large, number of detectors reduces but increases the voids. On the contrary, if the radius of the detector is small, detectors increase for the effective coverage of NS region. Figure 4.7 shows variable sized blue detectors around a single yellow self sample.

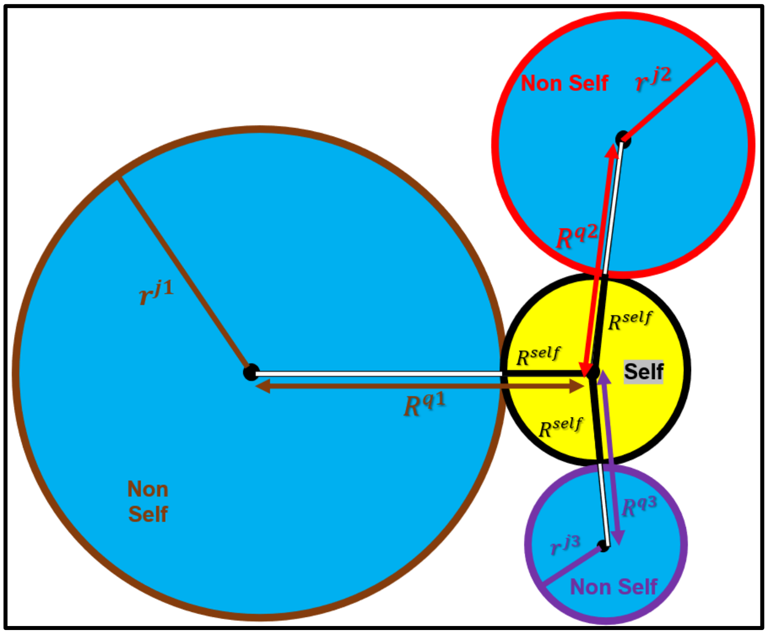


Figure 4.7. Variable sized NS detectors with a S sample in Vector Space

In this case, detector can be valid if equation (7) is satisfied

$$R^q > R^{self} \tag{7}$$

If any candidates fall within Self-region, its centre is adjusted by moving it away from training data samples. This may lead to overlapping of existing matures detectors, which is to be taken care of by moving candidate detectors away from matured detectors. In this way the distance between the center of detector and S-samples is maximized while the overlapping with matured detectors is minimized. This optimization of detector radius and minimization of detector overlap as shown in equation (8).

$$\max_{d_j} f_j^* = \max_{d_j} \frac{\delta f_j}{s_j} = \max_{d_j} \frac{\delta(|R^q| - R^k)}{s_j} \tag{8}$$

where f_j *is the initial fitness value and* f_j^* *is the updated fitness value,* δ *is the scaling factor for real valued detectors.*

$$s_j = \sum_{i=1}^{D} \xi_{ij} \tag{9}$$

$where\ \xi_{ij}\ is\ defined\ as\ distance\ based\ similarity\ between\ already$

$selected\ detacor\ and\ candidate\ detecotr.$

The diversity factor is given be s_j and ξ_{ij} is defined as similarity index based on the distance between the previously selected detectors and candidate detector. In this way if the new candidate detector is selected the oldest one with maximum similarity index is removed from the population. Following criteria can be used for maximization of detector radius r^j and minimizing the overlap among detectors as given in equation (10) to (11) [86].

$$\underset{d_j}{max}\ r^j = \underset{d_j}{max}|R^q| - R^k \tag{10}$$

where R^k is the radius of S sample.

Maximization of diversity factor

$$f_j^* = \frac{\delta f_j}{s_j}, \quad where\ s_j = \sum_{i=1}^{D} \xi_{ij} \tag{11}$$

The proposed scheme consists of calculation of distance-based similarity, calculation of initial fitness value, calculation of diversity factor and finally computation of final fitness value of detectors. The stopping criteria is number iterations set to a maximum value. The experiments have shown that at this maximum value of number of iterations, the change in fitness value function is very small.

4.3.2.2 Selection of Best Set of Optimized Detectors

The input given to GA for one of the class to calculate optimized detectors set on the basis of training data (S-samples). trailed by computation of detection accuracy using unknown test data (comprising of S and NS samples). The parametric combination for optimization of detectors and computation of detection accuracy comprises of compressed data set, number of detectors (ranging from 70-110 with a gap of 5) and R^{self} (ranging from 0.01-0.15 with a gap of 0.03). The detector set exhibiting the highest (best) detection accuracy is selected. The best selected detector set (i.e., Detector A for Class 0 and Detector B for Class 1) is used for classification of test sample after the detectors are optimized. Table 4.1 and Table 4.2 shows the detection accuracies for each detector with the radius of the S-elements i.e., Rself

Table 4.1 Detection accuracy of Detector A(Class 0)

Rself	**Detection Accuracy**	**Number of Detectors**
0.10	0.9000	85
0.15	**0.9143**	**85**
0.20	0.8429	80
0.25	0.7571	90
0.30	0.7000	75

Table 4.2 Detection accuracy of Detector B(Class 1)

rself	**Detection Accuracy**	**Number of Detectors**
0.10	**0.8429**	**100**
0.15	0.7857	90
0.20	0.7714	100
0.25	0.6571	90
0.30	0.6714	80

Our NSAIS Classifier comprises of two distinct detector sets, one for each class label, trained on S-samples (Training data) to predict the presence of heart disease. Table 4.3 shows the percentage of maximum training detection accuracies for both sets of detectors, which have been trained and selected to be used for classification.

Table 4.3 Max detection accuracy of the Detectors

Detector	Rself	Number of detectors	Max detection accuracy
A	**0.15**	**85**	**0.9143**
B	0.10	100	0.8429

These best detector sets for each class (A and B) are obtained through training (S) samples. During training of the classifier, individual training accuracy for prediction of absence of heart disease (Class 0) was 100 % and presence of heart disease (Class 1) was correctly predicted showing 94.44% training accuracy, as illustrated in the Table 4.4.

Table 4.4 Percentage training detection accuracies of selected detector sets

Training detection Accuracy (%)	
Class 0 **(No Heart Disease)**	**Class 1** **(Heart Disease Present)**
100.00	94.44

4.3.3 Training Model with Proposed BIC based on NSAIS Classifier

NSAIS Classifier is comprised of these two sets of detectors used for classification of heart disease. Test data is presented to NSAIS Classifier and decision of class is taken on the basis of affinity rule. The affinity of one test sample is checked for all detectors in a set. The affinity rule is given in equation (12)

$$dist(\omega^i, d^j) = \sqrt{\sum_{k=1}^{k}(\omega_k^i - d_k^j)^2} \tag{12}$$

where ω^i is the i^{th} test sample vector and d^j is the j^{th} detector and $dist$ is the Euclidian distance between the sample and detector. Each sample is checked for this affinity with all detectors of the set, if a match is established $(dist < threshold)$ between any of the detector and test sample, it is labeled as sample of non-self class. The threshold in our case is taken as the radius r^j of the detector which is being tested for affinity measure. The decision rule for class is given by equation (13)

$$if \begin{cases} dist + R^{self} < \ detector\ radius\ (sample\ does\ not\ belong\ to\ class) \\ \qquad\qquad belongs\ to\ class, otherwise \end{cases} \tag{13}$$

Our algorithm computes the classification accuracy for the test samples (unknown data) for both the classes. Workflow of BIC based NSAIS is shown in Figure 4.8.

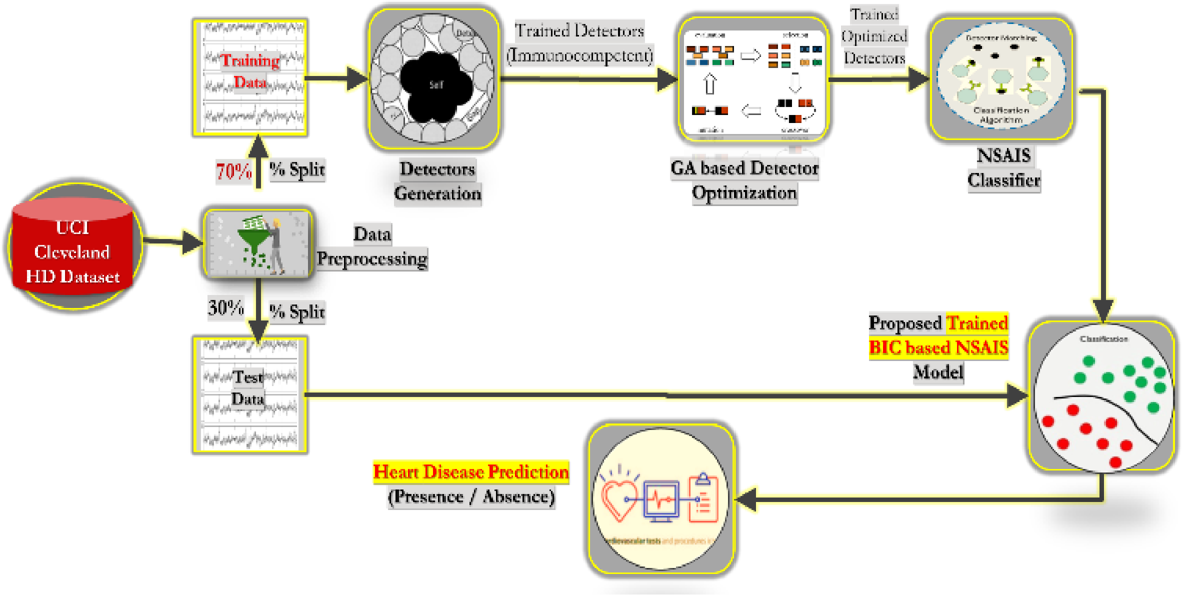

Training Data
Self
Trained Detectors (Immunocompetent)
Trained Optimized Detectors
70% % Split
Detectors Generation
GA based Detector Optimization
NSAIS Classifier
UCI Cleveland HD Dataset
Data Preprocessing
30% % Split
Proposed Trained BIC based NSAIS Model
Test Data
Classification
Heart Disease Prediction (Presence / Absence)

Figure 4.8. Flow diagram of proposed BIC based on NSAIS

4.3.4 Heart Disease Prediction through BIC based on NSAIS

The binary class data set used for classification of heart disease prediction shows, either the presence or absence of the heart disease. NSAIS Classifier is applied on the incoming training data and optimized trained detector sets. Two set of detectors (one of for each class) were thus obtained on the basis of detection accuracy. The detector set exhibiting best detection accuracy is selected. Theses optimized detector sets, computed for both the classes (i.e., detector A for class 0 and detector B for class 1), are used onwards for classification of test sample after the detectors are optimized. Our proposed trained BIC based NSAIS ML model also comprises of these two best detector sets trained on Self-samples (70 % Training Data). BIC based on NSAIS is a three-stage process:

(1) Detector generation and optimization;

(2) Selection of best set of the optimized detectors; and

(3) Classification of the test data set and computing the results

With 30% test data and confusion matrix, classification accuracy of the Proposed BIC based on NSAIS Model is computed to gauge the overall performance of a heart disease prediction system. The performance measure for BIC based on NSAIS is over all classification accuracy for binary classification can be calculated by equation (14).

$$Classification\ Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}\ x\ 100 \qquad (14)$$

Chapter 5

RESULTS

This chapter comprises of the results, and discussion on the research carried out during this study. A disease predcition system for precisely envisaging existence of heart disease, with UCI Cleveland Dataset, is designed through a novel BIC based on NSAIS approach, and optimized through GA. Comprehensive distinction of the proposed approach from the former work in literature, related to traditional ML Algorithms, has also been discussed. Results obtained are organized as under: -

- Heart disease prediction accuracy through
  - Traditional ML Algorithms (Mode-1)
  - Proposed BIC based on NSAIS (Mode-2)
- Comparison of two different data mining tools
- Comparison of results with various benchmark researches
- Traditional ML Algorithms vs Bio-inspired BIC based on NSAIS

5.1 Heart Disease Prediction

Improvement in accuracy for prediction of heart disease, has always remained a challenge. To compare the results, the proposed algorithm is applied on the same data set which has been used recently by other researchers. It was found that proposed BIC based on NSAIS approach (Mode-2) has shown significant improvement as compared to traditional ML methods (Mode-1) in the heart disease prediction accuracy.

5.1.1 Accuracy Results with Traditional ML Algorithms (Mode 1)

Five different traditional ML algorithms were applied to predict the classification accuracy for heart disease presence. Out of the five traditional classification algorithms used, Naïve Bayes performed better in Matlab for the highest classification accuracy for prediction of heart disease. Results are abridged in the Table 5.1.

Table 5.1 Classification accuracy (%) with Traditional ML Algorithms

	SVM	KNN	NB	Logistic Regression	Random Forest
Highest Prediction Accuracy	88.52	85.00	**94.36**	88.52	88.52

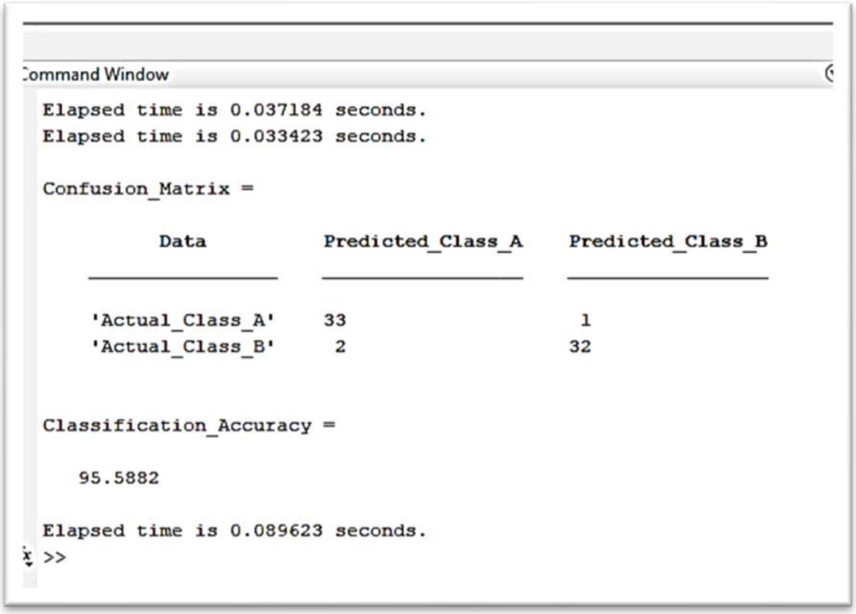
5.1.2 Accuracy Results with Proposed BIC based on NSAIS (Mode 2)

Once the test data was presented to our proposed BIC based NSAIS in Matlab environment, using 70% and 30% hold out validation. Performance measure for NSAIS i.e., Classification accuracy is calculated according to equation (14) depending upon the correct prediction of each class. Table 5.2 shows the classification accuracies using NSAIS.

Table 5.2 Percentage classification accuracy with BIC based on NSAIS

Patients/Instances	297 (6 Missing Values discarded)
Validation Type	Holdout Validation 70-30 % Split
Classifier	NSAIS with GA
Prediction Accuracy (%)	**95.58**

Heart disease prediction accuracy and the confusion matrix for the test sample data only, are shown in Figure 5.1.



```
Elapsed time is 0.037184 seconds.
Elapsed time is 0.033423 seconds.

Confusion_Matrix =

           Data           Predicted_Class_A    Predicted_Class_B
     ________________    _________________    _________________

     'Actual_Class_A'    33                    1
     'Actual_Class_B'     2                   32


Classification_Accuracy =

   95.5882

Elapsed time is 0.089623 seconds.
>>
```

Figure 5.1. Classification accuracy for BIC based on NSAIS

5.2 Analysis and Discussion

5.2.1 Classification Results in different Data Mining Tools

The accuracy results from two different ML environments i.e., Matlab & Weka were separately logged as well and then compared at the end. From the iterative results, it was clearly seen that the NB & LR gave more reliable results on Matlab. On the other hand, SVM & LR comparatively outperformed, when used in Weka environment. Highest classification accuracy of heart disease prediction remained with Naïve Bayes at 94.36% in Matlab & 88.52 % with RF, LR and SVM in Weka for the same dataset, as shown in Figure 5.2.

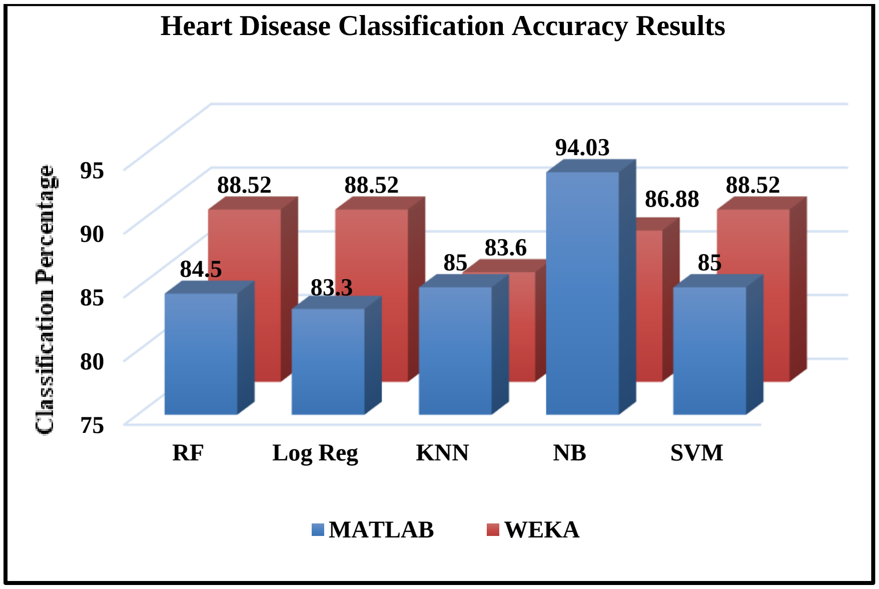


Figure 5.2. Classification Accuracies in two ML Platforms

5.2.2 Comparison with Various Benchmark Researches

BIC based on NSAIS has surpassed the classification accuracy from the bench mark research as shown in table 5.3. The accuracy results are statistically significant in comparison to all other traditional ML classifiers used in bench mark. Table 5.3 shows contrast between proposed approach and bench mark research.

Table 5.3 Comparison of results of proposed approaches with bench mark research

Approach	**Classification accuracy (%)**				
	Fahd S et al. (2019), [32],	**Al-Mustafa, et al. (2020),** [42]	**Nishanth Vaidya et al. 2020** [51]	**Proposed traditional ML Approach (Mode 1)**	**Proposed BIC based NSAIS Approach (Mode 2)**
	10-Fold Cross Validation **Decision Tree**	10-Fold Cross Validation **Decision Table**	Holdout 75-25% **SVM**	70-30 % Split Holdout Validation **Naïve Bayes**	70-30 % Split Holdout Validation **BIC based on NSAIS with GA**
Accuracy	93.19	93.85	94.60	94.03	**95.58**

Similarly, figure 5.3 shows the comparison of classification accuracies of proposed approaches with the other bench mark researches worldwide. BIC based on NSAIS has superseded all other classifiers used in bench mark approaches.

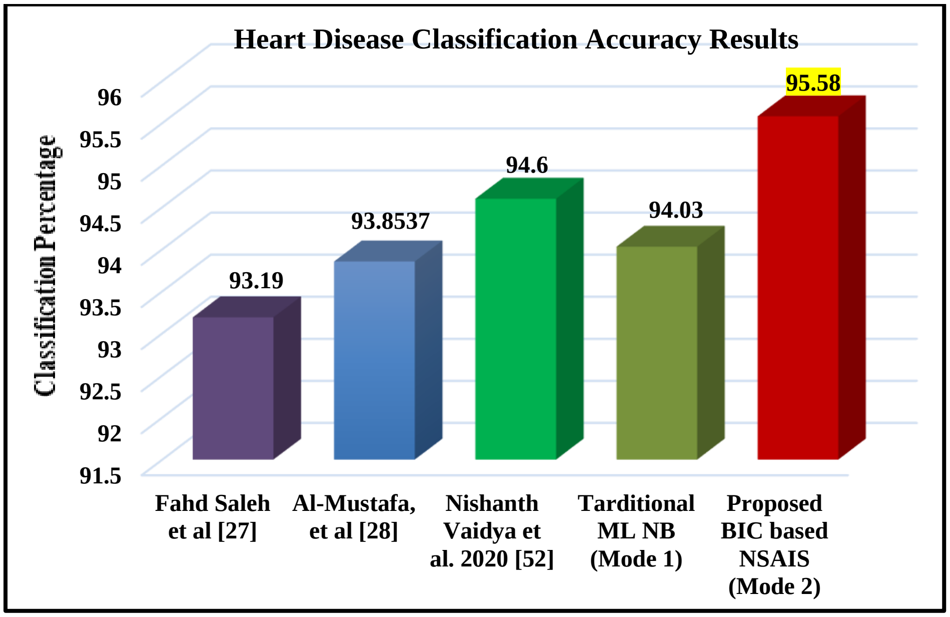


Figure 5.3.Accuracy with Proposed & Benchmark Approaches

5.2.3 Traditional ML Algorithms vs BIC based on NSAIS

Classification accuracy for prediction of heart disease with our proposed *Mode 2* i.e., BIC based NSAIS technique, was compared with *Mode 1* i.e., traditional ML classification techniques. Using the trained detectors, GA as a detector optimization technique and NSAIS as a classifier, our Bio-inspired ML technique in Mode 2, outperformed all the other traditional algorithms, as shown. BIC based on NSAIS has successfully for classified of presence of heart disease. Figure 5.4 shows the comparison of classification accuracy of our proposed approach in comparison with the traditional classification approaches.

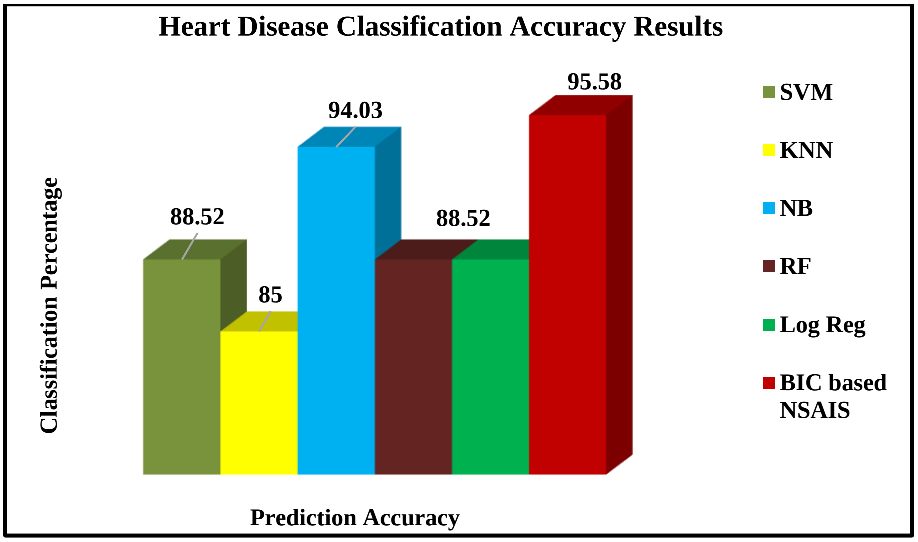


Figure 5.4. Accuracy with Traditional ML Algos vs BIC based on NSAIS

5.2.4 Performance Analysis: BIC based on NSAIS optimized with GA

. Our proposed approach (in Mode 2) with GA optimization, has outperformed the existing ML methods in our knowledge, by achieving maximum classification accuracy of **95.58%,** for the precise prediction of the heart disease. The optimization of detector radius using GA has increased performance of detection of NS by effectively covering the sample space. Hence, robustness and efficacy of using NSAIS as classifier coupled GA optimization has been proved by the improvement in classification accuracy over the known techniques mentioned in literature.

5.2.5 Discussion on Results

NSAIS Classifier has been effectively used for classification of heart disease presence, showing promising results in comparison to traditional classification techniques. As the classification accuracy not only depends upon the efficacy of classifier, but also upon the features selected in the dataset, variation in heart disease datasets obtained from different hospitals, needs to be catered.

5.3 Validation of the Results

In order to validate the performance of our proposed BIC based on NSAIS approach, miscellaneous datasets were obtained from multiple open-source data repositories.

5.3.1 Description – Miscellaneous Datasets

This data set was used in Kaggle [88], UCI Stat log [89], Enhanced UCI Cleveland [43] datasets.

5.3.2 Validation Results – Miscellaneous Datasets

Proposed methodology exhibited high classification accuracy when miscellaneous Datasets were checked with already trained BIC based on NSAIS classifier. Figure 5.7 shows the comparison of classification accuracies with different datasets. According proposed methodology of BIC based on NSAIS, two sets of optimized detectors are computed according to the number of classes. Table 5.4 shows number of detectors for maximum detection accuracy.

Table 5.4 Maximum detection accuracy

Class	Rself	Max detection accuracy	Number of detectors
0	0.15	**0.9143**	85
1	0.10	0.8429	100

Three set of detectors have been trained and optimized to classify another binary class heart disease data set. Table 5.5 shows the percentage of detection accuracies based on training for all sets of detectors.

Table 5.5 Percentage training detection accuracies

Dataset	Training detection Accuracy (%)	
	Class 0	Class 1
Kaggle [88]	100	94.44
UCI Stat log [89]		
Enhanced UCI Cleveland [43]		

Overall classification accuracy for binary class data for all the validation dataset used along with the number of patients is shown in the table 5.6and graphically depicted in figure 5.5.

Table 5.6 Percentage classification accuracy NSAIS

Dataset	Patients	Classification accuracy (%)
Kaggle [88]	1025	92.33
UCI Stat log [89]	270	98.33
Enhanced UCI Cleveland [43]	909	98.29
Mean	**~735**	**96.32**

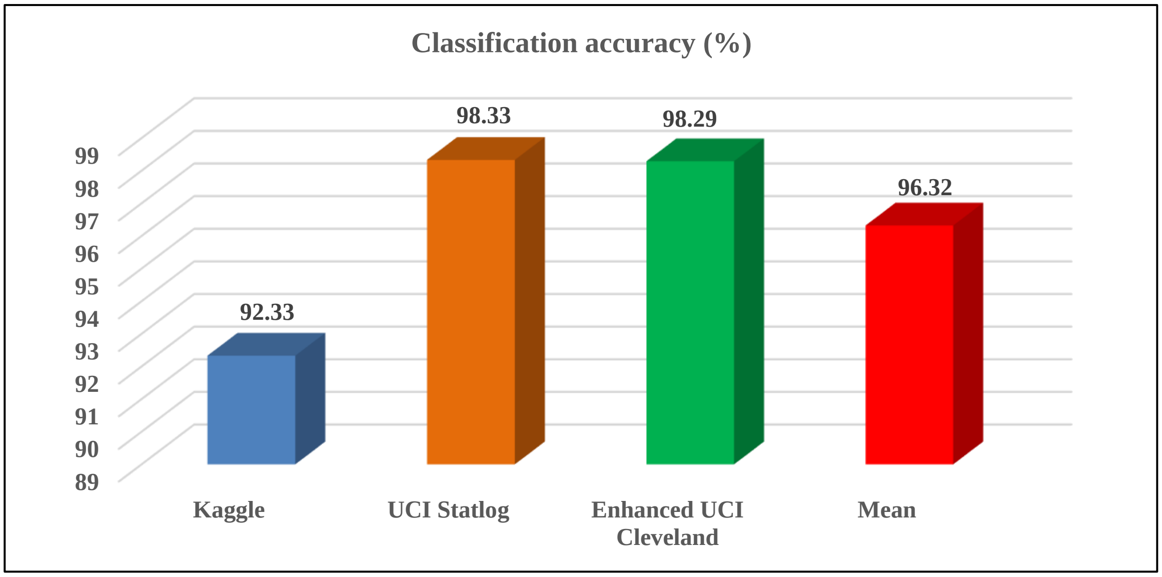


Figure 5.5 Percentage classification for Miscellaneous Validation Datasets

5.3.3 Analysis of Results - Miscellaneous Datasets

The results have shown that proposed Bio-inspired classification approach is exhibiting consistent higher classification accuracies, which proves robustness and efficacy of proposed algorithm on other open-source heart disease datasets. A mean classification accuracy using of 96.32% has been achieved by BIC based on NSAIS model using three different datasets.

Chapter 6

CONCLUSIONS AND RECOMMENDATIONS

This chapter consists of the contribution, conclusion and research guidelines for the future study. Keeping the research objectives in mind, the proposed Bio-inspired approach has been explored to develop an intelligent system to assist healthcare professionals in heart disease prediction. Proposed methodology for development of proposed BIC based NSAIS has been presented in chapter 4, while chapter 5 comprises of results, along with the comparison of results with bench mark and the validation of the results with miscellaneous datasets.

6.1 Contribution

The contribution of this thesis is to develop an intelligent predictive system to aid medical professionals in heart disease prognosis. Proposed approach is a bio-inspired innovative technique for classification. Although the NSAIS has already been used for applications of computer security and fault diagnosis, but to our knowledge, this has not been used as an independent classification technique for heart disease prediction as yet. The presented Bio-inspired classification arrangement, i.e., BIC based on NSAIS, comprises of two main phases: (1) Selection of optimized detector sets for each class using GA for detection of non-self samples using training data. (2) Using NSAIS (comprising of trained/optimized detector sets) as a classification method to predict presence of heart disease. The proposed

NSAIS Classifier has also been validated indicating highest classification accuracy in comparison to bench mark researches.

6.2 Conclusions

The proportion of heart disease patients is on an increase. To overcome such a perilous situation, there was a dire need of an intelligent system that can not only extract reliable information from the available hospital data, but also utilize it to train a classification system using ML methods. Henceforward, the traditional & Bio-inspired ML models were discussed in this study. It was found that little research has relatively focused on bio-inspired ML techniques for prediction of diseases. Thus, our research focused on BIC based on NSAIS framework as classification technique for the prediction of heart disease. In comparison with the previous researches, this study has shown significant improvement and higher accuracy results than previous works.

Literature review of available classifications methods for heart disease prediction was carried out and different ML techniques, showing high classification accuracies, have been discussed in chapter 2, which have led to set our bench mark classification accuracy. We used same UCI Cleveland data set for our study, which was used by the bench mark research for an eloquent comparison. The anatomy of heart, symptoms of heart disease and few concepts from BHIS and AIS were discussed in chapter 3. We also discussed the application of negative selection algorithm AIS (NSAIS) and focused on its application as a Bio-inspired classification technique for heart disease prediction which has not been applied in this domain according to our knowledge. Thus, making this technique a novel approach for predictive analysis of the heart disease. The framework of research methodology has been discussed in chapter 4 of the thesis, which is based upon classification using NSAIS. In the process, binary classification, steered us to select GA for the optimization of the detector set of each class and Bio-inspired classification method for classification of heart disease presence. In chapter 5, miscellaneous datasets were used for validation of proposed technique. The comparison was done on a standard performance evaluation matrix (i.e., accuracy).

6.3 Recommendations and Future Works

Future work can be expanded by using other data mining methods such as neural networks etc. along with using other untamed Bio-inspired Classification Techniques to calculate the classification accuracy for prediction of heart disease. Moreover, future research can also be taken up in ensuing:

- Application of NSAIS on other different types of patients with other allied diseases along with heart disease. Moreover, possibilities to be explored for detection of diseases like diabates, hepatitus, acute appendicitis, skin disease, epilepsy, breast cancer, liver disorders, hemorrhagic stroke & childhood leukemia etc.
- Development of a real-time web-based application or an online portal, can be designed in future, which can upload data directly from the apprehensive patients and the heart disease may be depicted using such innovative bio-inspired ML Approaches.
- Development of hybrid approach by combining NN and Fuzzy systems along with GA to predict the heart disease in real-time.

REFERENCES

[1] "World Health Organization - Noncommunicable Diseases (NCD) Country Profiles, 2018.".

[2] https://www.who.int/en/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[3] "Sharma, H., & Rizvi, M. A. (2017). Prediction of heart disease using machine learning algorithms: A survey. International Journal on Recent and Innovation Trends in Computing and Communication, 5(8), 99-104.".

[4] American Heart Association-Heart Disease and Stroke Statistics 2021 Update.

[5] Olaronke, I., & Oluwaseun, O. (2016, December). Big data in healthcare: Prospects, challenges and resolutions. In 2016 Future Technologies Conference (FTC) (pp. 1152-1157). IEEE..

[6] https://www.delltechnologies.com/en-ae/collaterals/unauth/infographic/solutions/dell-technologies-edge-iot-healthcare-infographic.pdf.

[7] Spencer, R., Thabtah, F., Abdelhamid, N., & Thompson, M. (2020). Exploring feature selection and classification methods for predicting heart disease. Digital health, 6, 2055207620914777..

[8] Jothi, N., & Husain, W. (2015). Data mining in healthcare–a review. Procedia computer science, 72, 306-313..

[9] Timmis, J., Neal, M., & Hunt, J. (2000). An artificial immune system for data analysis. Biosystems, 55(1-3), 143-150..

[10] "AK Pal, Rawal, P., Ruwala, R., & Patel, V. (2019). Generic disease prediction using symptoms with supervised machine learning. Int. J Sci. Res. Comput. Sci. Eng. Inf. Technol, 5(2), 1082-1086.".

[11] "Craig, W. J. (2009). Heart Healthy.".

[12] "https://my.clevelandclinic.org/health/diseases/16818-heart-attack-myocardial-infarction".

[13] "https://discover.hubpages.com/education/Anatomy-of-the-Heart".

[14] "Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal, 15, 104-116.".

[15] "Pushpalatha, S., & Pandya, D. (2016). Framework for Diagnosing Hepatitis Disease using Classification Algorithms. International Journal of Advanced Research, 4(7), 2189-2195.".

[16] "Akmese, O. F., Dogan, G., Kor, H., Erbay, H., & Demir, E. (2020). The Use of Machine Learning Approaches for the Diagnosis of Acute Appendicitis. Emergency medicine international, 2020.".

[17] "Verma, A. K., Pal, S., & Kumar, S. (2019). Classification of skin disease using ensemble data mining techniques. Asian Pacific journal of cancer prevention: APJCP, 20(6), 1887.".

[18] "Y Fan, & Chaovalitwongse, W. A. (2010). Optimizing feature selection to improve medical diagnosis. Annals of Operations Research, 174(1), 169-183.".

[19] "Wu, C. S. M., Badshah, M., & Bhagwat, V. (2019, July). Heart disease prediction using data mining techniques. In Proceedings of the 2019 2nd International Conference on Data Science and Information Technology (pp. 7-11).".

[20] "Lavanya, D., & Rani, D. K. U. (2011). Analysis of feature selection with classification: Breast cancer datasets. Indian Journal of Computer Science and Engineering (IJCSE), 2(5), 756-763.".

[21] "Bhardwaj, R., Mehta, R., & Ramani, P. (2020). A Comparative Study of Classification Algorithms for Predicting Liver Disorders. In Intelligent Computing Techniques for Smart Energy Systems (pp. 753-760). Springer, Singapore.".

[22] "Utami, E., & Raharjo, S. (2020, October). Mortality Prediction Using Data Mining Classification Techniques in Patients With Hemorrhagic Stroke. In 2020 8th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-5). IEEE.".

[23] "Labib, S. E., & Rayed, C. A. (2020). Prediction model for risk factors of childhood leukemia based on data mining classification algorithms. Egyptian Computer Science Journal, 44(2), 51-63.".

[24] "W. David, “Heart Disease Data Set,” Published by UCI, 1988. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.".

[25] "Tougui, I., Jilbab, A., & El Mhamdi, J. (2020). Heart disease classification using data mining tools and machine learning techniques. Health and Technology, 10, 1137-1144.".

[26] "Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2), 121-167".

[27] "Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern classification. John Wiley & Sons".

[28] "Ho, T. K. (1995, August). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.]".

[29] "[Georga, E. I., Tachos, N. S., Sakellarios, A. I., Kigka, V. I., Exarchos, T. P., Pelosi, G., ... & Fotiadis, D. I. (2019). Artificial intelligence and data mining methods for cardiovascular risk prediction. In Cardiovascular Computing—Methodologies and C".

[30] "David. H, Benjamin & Belcy, S.. (2018). HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES. 09. 1824-1830. 10.21917/ijsc.2018.0254.".

[31] "Mary, T. S., & Sebastian, S. (2019). Predicting heart ailment in patients with varying number of features using data mining techniques. International Journal of Electrical and Computer Engineering, 9(4), 2675.".

[32] "Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. International Journal of Advanced Computer Science and Applications, 10(6), 261-268.]".

[33] "J. Soni, “Predictive Data Mining for Medical Diagnosis : An Overview of Heart Disease Prediction,” vol. 17, no. 8, pp. 43–48, 2011".

[34] "Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In 2019 16th international bhurban conference on applied sciences and technology (IBCAST) (pp. 619-623).".

[35] "Panda, D., & Dash, S. R. (2020). Predictive system: Comparison of classification techniques for effective prediction of heart disease. In Smart intelligent computing and applications (pp. 203-213). Springer, Singapore.".

[36] "Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203.]".

[37] "Gultepe, Y., & Rashed, S. (2019). The use of data mining techniques in heart disease prediction. Int. J. Comput. Sci. Mob. Comput, 8(4), 136-141.".

[38] "[Khodke, H. E., Yadav, S. K., & Kyatanav, D. N. (2020, October). A new Approach of Heart Disease Prediction system using Data Science. In 2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing (ICSIDE".

[39] "[Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Information Systems, 2018.]".

[40] "[Emakhu, J., Shrestha, S., & Arslanturk, S. Prediction System for Heart Disease Based on Ensemble Classifiers.]".

[41] "[Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., & Ghuli, P. Heart Disease Prediction using Machine Learning. April-2020]".

[42] "Almustafa, K. M. (2020). Prediction of heart disease and classifiers' sensitivity analysis. BMC bioinformatics, 21(1), 1-18".

[43] "[Nidhi Bhatla, kiryan Joti_An Analysis of Heart Disease Prediction using Different Datamining techs]".

[44] "Swain, D., Ballal, P., Dolase, V., Dash, B., & Santhappan, J. (2020). An Efficient Heart Disease Prediction System Using Machine Learning. In Machine Learning and Information Processing (pp. 39-50). Springer, Singapore.".

[45] "Marimuthu, M., Abinaya, M., Hariesh, K. S., Madhankumar, K., & Pavithra, V. (2018). A review on heart disease prediction using machine learning and data analytics approach. International Journal of Computer Applications, 181(18), 20-25.".

[46] "Kanchan, B. D., & Kishor, M. M. (2016, December). Study of machine learning algorithms for special disease prediction using principal of component analysis. In 2016 international conference on global trends in signal processing, information computing".

[47] "Ekız, S., & Erdoğmuş, P. (2017, April). Comparative study of heart disease classification. In 2017 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4). IEEE.".

[48] "El-Bialy, R., Salamay, M. A., Karam, O. H., & Khalifa, M. E. (2015). Feature analysis of coronary artery heart disease data sets. Procedia Computer Science, 65, 459-468".

[49] "Alizadehsani, R., Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Panahiazar, M, & Sarrafzadegan, N. (2019). A database for using machine learning and data mining techniques for coronary artery disease diagnosis. Scientific data, 6(1), 1-13.," in *www.cadataset.com*.

[50] "Jagtap, A., Malewadkar, P., Baswat, O., & Rambade, H. (2019). Heart disease prediction using machine learning. International Journal of Research in Engineering, Science and Management, 2(2), 352-355.".

[51] "Vaidya, N., Kandu, M., Yadav, R. K., & Bharadwaj, N. (2020). The Working of Various Prediction Techniques For Heart Diseases–A Case Study. IJRAR-International Journal of Research and Analytical Reviews (IJRAR), 7(1), 447-453".

[52] "Abid, A., Khan, M. T., Haq, I. U., Anwar, S., & Iqbal, J. (2020). An improved negative selection algorithm-based fault detection method. IETE Journal of Research, 1-12.".

[53] "Forrest, S., Perelson, A. S., Allen, L., & Cherukuri, R. (1994, May). Self-nonself discrimination in a computer. In Proceedings of 1994 IEEE computer society symposium on research in security and privacy (pp. 202-212).".

[54] "Guerroumi, M., & Derhab, A. (2019). NSNAD: negative selection-based network anomaly detection approach with relevant feature subset. Neural Computing and Applications, 1-27.".

[55] "Luther, K., Bye, R., Alpcan, T., Muller, A., & Albayrak, S. (2007, June). A cooperative AIS framework for intrusion detection. In 2007 IEEE International Conference on Communications (pp. 1409-1416). IEEE.".

[56] "Anaya, M., Tibaduiza, D. A., & Pozo, F. (2015). A bioinspired methodology based on an artificial immune system for damage detection in structural health monitoring. Shock and Vibration, 2015".

[57] "Li, D., Liu, S., Gao, F., & Sun, X. (2020). Continual learning classification method with new labeled data based on the artificial immune system. Applied Soft Computing, 94, 106423.".

[58] "Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464-1480.".

[59] "Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. IEEE Transactions on neural networks, 3(5), 698-713.".

[60] "Hoya, T., Hori, G., Bakardjian, H., Nishimura, T., Suzuki, T., Miyawaki, Y., ... & Cao, J. (2003, January). Classification of single trial EEG signals by a combined principal+ independent component analysis and probabilistic neural network approach. In Pr".

[61] "Penny, W. D., Roberts, S. J., Curran, E. A., & Stokes, M. J. (2000). EEG-based communication: a pattern recognition approach. IEEE transactions on Rehabilitation Engineering, 8(2), 214-215.".

[62] "Kostov, A., & Polak, M. (2000). Parallel man-machine training in development of EEG-based cursor control. IEEE Transactions on Rehabilitation Engineering, 8(2), 203-205.".

[63] "Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1-2), 82-89.".

[64] "Pasha, A., & Latha, P. H. (2020). Bio-inspired dimensionality reduction for Parkinson’s disease (PD) classification. Health information science and systems, 8(1), 1-22.".

[65] "Rashid, N., Iqbal, J., Mahmood, F., Abid, A., Khan, U. S., & Tiwana, M. I. (2018). Artificial immune system–negative selection classification algorithm (NSCA) for four class electroencephalogram (EEG) signals. Frontiers in human neuroscience, 12, 439.".

[66] "Ghorbani, R., & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. International Journal of Data and Network Science, 3(2), 47-70".

[67] "Akella, A., & Kaushik, V. (2020). Machine Learning Algorithms for Predicting Coronary Artery Disease: Efforts Toward an Open Source Solution. bioRxiv".

[68] "Kahramanli, H., & Allahverdi, N. (2008). Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1-2), 82-89.".

[69] "Kemal Polat, Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing".

[70] "Sharma, M., Bansal, A., Gupta, S., Asija, C., & Deswal, S. (2020). Bio-inspired algorithms for diagnosis of heart disease. In International conference on innovative computing and communications (pp. 531-542). Springer, Singapore.".

[71] "Vijayashree, J., & Sultana, H. P. (2018). A machine learning framework for feature selection in heart disease classification using improved particle swarm optimization with support vector machine classifier. Programming and Computer Software, 44(6), 388-3".

[72] "Sharma, S., & Parmar, M. (2020). Heart Diseases Prediction using Deep Learning Neural Network Model. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9(3)".

[73] Abbas, A. K., Lichtman, A. H., & Pillai, S. (2014). Basic immunology: functions and disorders of the immune system. Elsevier Health Sciences..

[74] "Fischer, J. C., Zänker, K., Van Griensven, M., Schneider, M., Kindgen-Milles, D., Knoefel, W. T & Matuschek, C. (2020). The role of passive immunization in the age of SARS-CoV-2: an update. European journal of medical research, 25, 1-6.".

[75] "De Castro, L. N., & Von Zuben, F. J. (1999). Artificial immune systems: Part I–basic theory and applications. Universidade Estadual de Campinas, Dezembro de, Tech. Rep, 210(1).".

[76] "de Castro, L. N., & Timmis, J. I. (2003). Artificial immune systems as a novel soft computing paradigm. Soft computing, 7(8), 526-544.".

[77] "Tonegawa, S. (1983). Somatic generation of antibody diversity. Nature, 302(5909), 575-581.".

[78] "Nossal, G. J. V. (1994). Negative selection of lymphocytes. cell, 76(2), 229-239.".

[79] "Mannie, M. D. (1999). Immunological self/nonself discrimination. Immunologic research, 19(1), 65-87.".

[80] "Zheng, J., Chen, Y., & Zhang, W. (2010). A survey of artificial immune applications. Artificial Intelligence Review, 34(1), 19-34.".

[81] "M. Mitchell. An introduction to genetic algorithms. MIT press, 1998.".

[82] "Urso, A., Fiannaca, A., La Rosa, M., & Rizzo, R. (2018). Data mining Prediction methods. Encycl. Bioinforma. Comput. Biol. ABC Bioinforma, 413.".

[83] "Watkins, A., & Timmis, J. (2004, September). Exploiting parallelism inherent in AIRS, an artificial immune classifier. In International Conference on Artificial Immune Systems (pp. 427-438). Springer, Berlin, Heidelberg.".

[84] "Hart, E., & Timmis, J. (2008). Application areas of AIS: The past, the present and the future. Applied soft computing, 8(1), 191-201.".

[85] "Lau, H. Y., Wong, V. W., & Lee, I. S. (2007). Immunity-based autonomous guided vehicles control. Applied Soft Computing, 7(1), 41-57.".

[86] "Abid, A., Khan, M. T., & Khan, M. S. (2017). Multidomain Features-Based GA Optimized Artificial Immune System for Bearing Fault Detection. IEEE Transactions on Systems, Man, and Cybernetics: Systems.".

[87] "Ji, Z., & Dasgupta, D. (2004, June). Real-valued negative selection algorithm with variable-sized detectors. In Genetic and Evolutionary Computation Conference (pp. 287-298). Springer, Berlin, Heidelberg.".

[88] "Heart Disease Dataset. (2019, June 6). Kaggle. https://www.kaggle.com/johnsmith88/heart-disease-dataset".

[89] " Statlog database: http://archive.ics.uci.edu/ml/ machine- learning-databases/statlog/heart".

[90] "Alizadehsani, R.Roshanzamir, M., Abdar, M., Beykikhoshk, A., Khosravi, A., Panahiazar, M., ... & Sarrafzadegan, N. (2019). A database for using machine learning and data mining techniques".